Open Research Statement:

Data and MATLAB codes for results reported here are available in the Github repository

https://anonymous.4open.science/r/MATLAB-reconstruction-package-3716/README.md

Abstract

1. The ecological literature often features phenomenological dynamic models lacking robust validation against observational data. Reverse engineering is an alternative approach, where time series data are utilized to infer or fit a stochastic differential equation. This process, known as system reconstruction, presents significant challenges especially when data resolution is low. This paper addresses the estimation of the (often) non-linear deterministic and stochastic parts of Langevin models for sparsely sampled time series.

2. We introduce a Maximum Likelihood Estimation (MLE) inference method, termed Euler reconstruction, tailored for time series data with medium to high resolution. However, the Euler approach is not reliable for low-resolution data. To fill the gap for sparsely sampled data, we present an MLE inference method pioneered by Aït-Sahalia, that we term Hermite reconstruction. We employ a powerful modeling framework utilizing splines to detect inherent nonlinearities in the unknown data-generating system to achieve high accuracy with minimal computational burden.

3. We applied both methods to a range of simulated, ecological, and climate datasets, with different data resolutions. We provide a practical measure ('relaxation time') to distinguish between different data resolutions. We show that the Euler reconstruction can accurately reveal the underlying system when the data resolution is medium to high. The Hermite reconstruction can even recover models with low data resolution.

4. Our proposed inference technique for stochastic Langevin systems excels in revealing the nonlinearities of the data-generating system even when data resolution is low. This is mainly due to the utilization of spline modeling. The provision of a MATLAB package and a user-friendly tutorial further facilitates researchers in the life sciences to reconstruct their own data easily and conveniently.

KEYWORDS

Langevin models, system reconstruction, Data resolution, Euler reconstruction, Hermite reconstruction, spline modeling

Introduction. It has long been a matter of debate whether ecosystems can have alternative stable states, how to measure their resilience and whether they can recover from perturbations. These are fundamental ecological questions which have mainly been discussed using theoretical models (Connell & Sousa 1983). Tackling these questions necessitates a reverse engineering approach, where we reconstruct the system based on dynamic data (Siegert & Friedrich 2001; Rinn *et al.* 2016). Subsequently, the resilience of the best-fitting model can be studied (Bolker *et al.* 2013; Hilborn & Mangel 2013; Arani *et al.* 2021). When appropriate mechanistic dynamical equations are known, parameters of the equations can be inferred from time series data.

Ecologists are often confronted with a situation where the nonlinear stochastic model that generated the data is unknown and must be inferred from data. If data are measured frequently throughout the entire range of the state variables, then a nonlinear stochastic model can be reconstructed by nonparametric methods (Bandi & Phillips 2003; Rinn *et al.* 2016). The reconstructed model can then be analyzed to map the stability landscape, locate stable or unstable equilibria, and calculate stochastic indicators of resilience such as mean exit time or median survival time (Arani *et al.* 2021). While nonparametric reconstruction methods are powerful, they have several limitations. They require high-resolution data and involve arbitrary choices, such as estimating the conditional mean, conditional variance, and higher conditional moments using kernels with bandwidths that control smoothness. Estimating the bandwidth is challenging, and the results

may be sensitive to its choice. Additionally, these methods extrapolate moments to zero for a specified first few numbers of lags, defined as integer multiples of the time step in the data (Siegert & Friedrich 2001; Bandi & Phillips 2003; Rinn *et al*. 2016). The choice of the number of lags significantly affects the results, particularly when data resolution is low. Furthermore, these methods demand substantial amounts of data for reliable estimation (Siegert et al., 1998; Rinn et al., 2016). An alternative approach involves model-fitting techniques that maximize a likelihood function. This approach includes both the classical inference scheme and a novel methodology pioneered by Aït-Sahalia (Aït-Sahalia 2002). It offers an attractive method that requires fewer data for reliable parameter estimation.

This paper presents a maximum likelihood estimation (MLE) approach for reconstructing the following one-dimensional Langevin model from time series data

$$dx = \mu(x;\theta)dt + \sigma(x;\theta)dW, \tag{1}$$

where $\mu(x;\theta)$ denotes the deterministic component of the system, referred to as the 'drift function' and $\sigma(x;\theta)$ represents the stochastic component, known as the 'diffusion function'. θ represents the parameter vector to be estimated via MLE and W refers to a Wiener process, under which the noise source dW, known as Brownian noise, is Gaussian and white (uncorrelated). The diffusion function $\sigma(x;\theta)$ weighs the impact of noise, reflecting the intensity of perturbations at state x. The noise in a diffusion model (1~~+~~*1*) is additive if the diffusion function does not depend on the state x (i.e., it is constant in terms of state), otherwise, it is called multiplicative. The inference technique employed here is parametric meaning that a parametric form, with a single vector of parameters θ, for both the drift $\mu(x;\theta)$ and diffusion $\sigma(x;\theta)$ functions should be predefined and then the parameters are estimated using time series data through MLE.

Formatted: Font: Not Italic

In this study, we apply the classical inference of Langevin models using Euler methodology which we call 'Euler reconstruction'. This method requires data of medium or high resolution, posing a significant challenge in fields such as ecology and climate science. To address this issue, we present an MLE scheme based on the groundbreaking work of Aït-Sahalia (Aït-Sahalia 2002) for univariate sparsely sampled data.

The idea of Aït-Sahalia involves constructing an approximate conditional density, and therefore an approximate MLE or quasi MLE, of (11) using Hermite polynomials, so we call this procedure 'Hermite reconstruction'. To our knowledge, this paper is the first to provide an open-access MATLAB package and to demonstrate successful applications to ecosystem and climate data. One of our key techniques for addressing the underlying MLE problem is the utilization of a wide variety of spline models (De Boor & De Boor 1978) as representation of the unknown drift and diffusion functions. Splines are flexible non-linear structures in terms of the state variable x, yet they are linear functions of the parameter vector θ, significantly enhancing the accuracy and speed of calculations (as demonstrated in various examples in our user-friendly tutorial and here. Technical details in Method's section). Additionally, splines serve as valuable tools in cases where selecting an appropriate model is uncertain, which is often encountered. To distinguish this modeling technique from typical 'parametric modeling' we term it 'spline modeling'.

The remainder of the paper presents the steps of the analysis, demonstrates reconstruction of diffusion models for several simulated, ecological, and climate data and shows that the method succeeds even for rarified datasets. Supplementary materials include a mathematical appendix, and a tutorial to illustrating a broader range of the capabilities of the Euler and Hermite reconstructions using both parametric and spline models across different data types. These data types encompass typical time series data (i.e., single time series), replicate time series (multiple time series belonging to the same data-generating system), large datasets (where a well-mixed small fraction of data will be used instead of the entire dataset), datasets with missing values, or various combinations of these.

Steps of the Method.

Step 1. Prepare the data: data standardization

In some real datasets, the scale (range) of the data can be very large. In such cases, data standardization, achieved by computing z-scores (subtracting the mean and dividing by the standard deviation), makes it easier to solve the MLE. This is because standardization helps to narrow down large search spaces, making

Formatted: Font: Not Italic

them more manageable. Additionally, standardization brings the data to a common scale, centered at 0 with small dispersion. This, in turn, makes it convenient to define a small region of parameter space for the MLE algorithm to search within especially for spline modeling (See Step4 for more details on this). For further discussion on data standardization, refer to section 8 of tutorial and Appendix G if you are interested in technical details.

Step 2. Check the data requirements

There are three data requirements for a Langevin model which should be checked prior to performing the reconstruction (for a detailed discussion see section 6 of the tutorial). Firstly, data should be stationary, meaning that its statistical properties remain unchanged throughout the study period. If this assumption is violated, the data should be divided into smaller (possibly overlapping) periods where stationarity is assured. Reconstruction can then be carried out separately for each period, with the final system reconstruction obtained by interpolating the results. Secondly, data should be Markovian meaning that the future state of data, given the present state, should be independent of the entire past history of states. The smallest time scale at which Markovicity holds is called the 'Markov-Einstein' (ME) time scale. Reconstruction can, then, be performed on a rarified sample of data whose resolution matches the (ME) time scale (Friedrich et al. 2011) or on samples with lower resolutions. To estimate ME time scale, we fit an autoregressive (AR) model to the data and examine its order, say p. If the fitted AR model has order one ($p = 1$), this strongly suggests that the ME time scale equals the data's sampling time, indicating that the dataset is Markovian. Conversely, if the AR model has an order greater than one ($p > 1$), the ME time scale is p times the data's sampling time, meaning the dataset is not inherently Markovian. However, a rarified sample, taken every p^{th} data point, will exhibit Markovicity, and reconstruction should be performed on this rarified sample. Thirdly, it is essential to check the resolution of the rarified sample regarded now as our dataset to be analyzed. This is the subject of next step.

Step 3. Check the resolution of data: how high should the data resolution be?

How high should the data resolution be in order to be able to reconstruct the data-generating system? This is a question we should investigate prior to performing any reconstruction procedure. The answer depends on the 'speed' or 'time scale' of the yet unknown system relative to sampling frequency of the data. To investigate the resolution of data the autocorrelation function of data should be examined and a quantity known as '*relaxation time*' τ_R should be estimated. As its name indicates, relaxation time is the time it takes for a process to get 'relaxed' or become 'statistically independent' of the starting state. We can roughly estimate relaxation time directly from data by fitting the exponential $\exp(-ct)$ to the initial lags of the data autocorrelation function, obtaining the estimate $\hat{c}$, and the relaxation time $\tau_R = 1/\hat{c}$ (Honisch *et al.* 2012) (refer to Appendix A for the details). Assuming Δ to be the data sampling time we consider the following three cases.

The first case: Δ ***being much smaller than*** τ_R $(\Delta \ll \tau_R)$

In this high-resolution regime we can safely apply simple reconstruction schemes like Euler scheme or Langevin approach (Rinn *et al.* 2016) (see Figure 1). Unfortunately, many real datasets are not sampled frequently enough to fall into this regime. Furthermore, some high-resolution data do not adhere to the Markov property, which is another data requirement (see Step 3). However, a rarified sample of such data may be Markovian but this comes at the expense of reduced resolution.

The second case: Δ ***being approximately the same order of magnitude as*** τ_R

In theory, it may be possible to reconstruct the underlying system in this case using more sophisticated reconstruction algorithms (Anteneodo & Queirós 2010; Honisch *et al.* 2012). In practice, however, current reconstruction procedures typically perform reliably when the sampling time does not exceed the relaxation time, i.e., $\Delta \leq \tau_R$. Consequently, in this work we consider $\Delta = \tau_R$ as the minimum resolution required for accurate data reconstruction. This resolution limit defines a critical resolution in practice (see Figure 1). In such cases (i.e., Δ not being small but still less than or equal to τ_R), a more accurate reconstruction procedure than the Euler scheme is often necessary. Here, Hermite reconstruction becomes essential (see Figure 1).

Refer to Example 5, which showcases a dataset accurately reconstructed with high precision using Hermite reconstruction at a resolution matching τ_R.

***The third case:* Δ *being much bigger than* τ_R ($\Delta \gg \tau_R$)**

In this case, consecutive measurements are almost independent. Therefore, the true dynamics is not reflected in such datasets, and reconstruction procedures are likely to fail in this 'independence limit' (Anteneodo & Queirós 2010) due to systematic errors termed 'finite time effects' in the literature (Honisch *et al.* 2012) (see Figure 1). Increasing the dataset size (i.e., the number of data points) does not resolve this issue—even a very large dataset remains ineffective.

Here, we provide a convention for categorizing data resolution into three regimes when $\Delta \leq \tau_R$: low, medium, and high. In particular, we say that a dataset has a low-resolution if Δ falls in the interval $(0, 0.01\tau_R]$, has a medium-resolution if Δ falls in the interval $(0.01\tau_R, 0.02\ \tau_R]$, and finally has a low-resolution if Δ falls in the interval $(0.02\tau_R,\ \tau_R]$. While Euler reconstruction is suitable for high- and medium-resolution regimes, Hermite reconstruction may be required for the low-resolution regime (see Figure 1).

The above-mentioned convention is an 'empirical' guideline based on our experience applying Euler and Hermite reconstructions to a wide range of simulated and real datasets (details in Step 5 and Step 6). It is important to note that these regimes of resolutions provide only an 'approximate' picture, as the relaxation time τ_R of the yet unknown data-generating system can only be estimated. Thus, this convention offers a general sense of the data resolution rather than a precise calculation

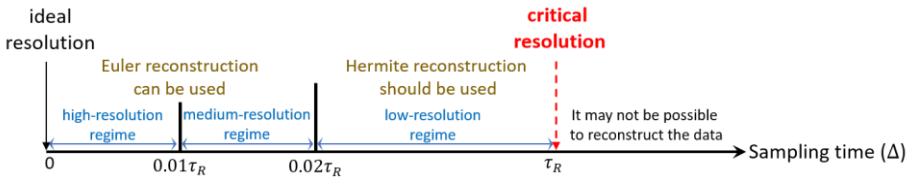

Figure 1. A schematic illustration of different regimes of resolution, along with appropriate reconstruction procedures for each regime. High-resolution, medium-resolution, and low-resolution data correspond to time scales within the intervals $(0, 0.01\tau_R]$, $(0.01\tau_R, 0.02\tau_R]$, and $(0.02\tau_R,\ \tau_R)$, respectively, where τ_R represents the data

relaxation time and can be directly estimated from the correlational information in the data. Euler reconstruction is suitable for data with high and medium resolution, whereas Hermite reconstruction is required for low-resolution data. If the data sampling time exceeds the relaxation time, reconstruction of such a poor dataset may be impossible.

Step 4. Guidelines for implementing parametric and spline modeling

In parametric modeling, parametric forms for the drift and diffusion functions in the Langevin model (1~~+~~*11*) should be specified prior to embarking on MLE. For instance, in the Ornstein-Uhlenbeck model the parametric forms for the drift and diffusion functions are $\mu(x) = -\mu x$ and $\sigma(x) = \sigma$, respectively where $\theta = [\mu, \sigma]$ is the vector of parameters. Similarly, in the grazing model of May (May 1977) $\mu(x) = rx\left(1 - \frac{x}{K}\right) - \frac{\gamma x^2}{x^2+a^2}$, $\sigma(x) = \sigma$ where x represents the biomass of a plant population, $\theta = [r, K, \gamma, a, \sigma]$ is the vector of parameters in which, r is the growth rate, K is the carrying capacity, γ is the maximum grazing rate, a is the efficiency of the grazer, and σ is the intensity of environmental perturbations which is assumed to be constant (so, this is an additive model).

In spline modeling (De Boor & De Boor 1978), on the other hand, no specific model for drift and diffusion functions needs to be specified in the conventional sense. Instead, a rather sparse mesh across the data range, called 'knot sequence', is specified. Knots are the x-coordinates of points where the spline interpolation goes through and the values of the drift and diffusion spline functions at these knots serve as model parameters (see red stars in Figures 2-9). Simply put, splines are piecewise polynomials that are smoothly joined to construct a 'complex' functional form. Therefore, splines are 'flexible' structures which enable us to recover the unknown nonlinearities of drift function $\mu(x)$ and diffusion function $\sigma(x)$. Nonetheless, splines are 'simple' structures because they are composed of polynomial pieces. We often use cubic splines, especially for Euler reconstruction. However, when dealing with Hermite reconstruction of highly low-resolution data we employ quadratic splines, as they significantly reduce the computational burden (see Appendix G for details. Also, see section 10 in the tutorial).

Formatted: Font: Not Italic

We select equidistant knots to minimize the sensitivity of the fitted spline to knot values, i.e., model parameters (De Boor & De Boor 1978) (details in Appendix G). Moreover, using a large number of knots is generally not advisable for several reasons. First, a large number of knots increases the complexity of the MLE problem, leading to numerous local maxima, and making the problem both challenging and time-consuming to solve. Second, in many practical scenarios datasets are small, and using a large number of knots can produce bumpy estimates for the drift and diffusion functions. Third, we often encounter unimodal or bimodal datasets, where the underlying system has either one equilibrium or three equilibria (two stable and one in the middle being unstable). In such cases, four knots are sufficient to estimate drift and diffusion functions qualitatively. For reasonably accurate quantitative results, eight knots are typically sufficient, as employed in many of our examples. However, in two cases (Example 5 and Example 8) with low-resolution data, we used seven knots to speed up parameters estimation in Hermite reconstruction.

After selecting a model (parametric or spline), a vector of lower bounds and a vector of upper bounds for the parameters should be specified for the optimization (MLE) to search within and find the optimal parameter values. All parameters associated with the diffusion function should be bounded in a way that ensures the diffusion function remains positive. For parametric models, the physics of the problem can provide insights to determine proper vectors of lower and upper bounds. For example, in the case of the May model, we know that all model parameters should be positive as they represent ecological quantities. For spline models selecting vectors of lower and upper bounds are convenient since we deal with stationary data. For instance, for the drift parameters, we can set all lower bounds to $-L$ and all upper bounds to L while for diffusion parameters we can set all lower bounds to 0 and all upper bounds to L where $L > 0$ should be big enough. In all the examples in this paper, we set $L = 10$ (although $L = 5$ was also sufficient). It is straightforward to verify whether the chosen L is sufficiently large: if any of the model parameters approaches L closely, it suggests that L needs to be increased.

If a parametric model is preferred and it is uncertain to choose a proper model then it is advisable to select drift and diffusion models which contain constant, linear, quadratic, and higher order terms due to Taylor series expansion. An additive version of such a modeling is

$$dx = \{\alpha + Ax + Bx^2 + \ldots\}dt + \sigma dW\,, \tag{2}$$

The parameter A is particularly insightful as it is the sole determinant of stability of the deterministic part of (~~2~~2). The model form in (~~2~~2) is not directly suited for fitting to data. For convenience and ease of parameter estimation, data should be standardized first, and then the model (~~2~~2) could be applied. Since, this model is linear in parameters, it is convenient to transform the estimated parameters back into their original scales by replacing the state variable x with its standardization $(x - m)/s$ and multiplying the right-hand side by s where m and s are the mean and standard deviation of the data, respectively.

Step 5. Euler reconstruction

If the data resolution falls within either the high-resolution or medium-resolution regimes, the general Langevin model (~~1~~1) can be approximated by a difference equation using the Euler-Maruyama discretization scheme (Gardiner 1985). This approach allows us to derive an approximate closed form for the conditional distribution of model (~~1~~1), which is typically unavailable and represents the primary challenge in estimating the parameters of Langevin models. The closed form, in turn, enables the construction of an approximate likelihood function (quasi-MLE). From this point, parameter estimation becomes straightforward: we solve the quasi-MLE as an optimization problem. For technical details, interested readers are referred to Appendix D.

Step 6. Hermite reconstruction

If the data resolution falls in the category of low-resolution regime, Hermite reconstruction should be employed. Here, we briefly outline the approach, but for detailed mathematical explanation, refer to Appendices F, G, H and I. For a less technical explanation refer to section 11 of the tutorial, which includes

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Formatted: Font: Not Italic

Formatted: Font: Not Italic

numerous case studies. In our MATLAB package, we have implemented the Hermite reconstruction based on a refinement by (Bakshi & Ju 2005) to a methodology developed by (Aït-Sahalia 2002).

Our approach involves a two-phase algorithm. In the first phase Euler reconstruction is performed, resulting in a first guess of the parameter vector called $\hat{\theta}_{Euler}$. In the second phase, the algorithm improves this first guess using Hermite reconstruction.

Reconstructing datasets with lower resolution becomes computationally more intensive. The primary challenge lies in dealing with an optimization problem with a partially-defined objective function in the second phase. As data resolution diminishes, the algorithm should search within smaller regions in the parameter space. To address this challenge, the algorithm first identifies a set of parameters, we call 'legitimate points' (LP), where the Hermite objective function has defined values. Subsequently, the algorithm utilizes these LPs and searches in the vicinity of the Euler estimation $\hat{\theta}_{Euler}$ to get an improved vector of parameters $\hat{\theta}_{Hermite}$. Refer to Appendix G for details of the optimization process.

Our package not only supports commonly used cubic splines but also offers a range of other spline types. The use of 'quadratic' splines is particularly significant for achieving considerably faster speeds compared to cubic splines in Hermite reconstruction. Finally, while both parametric and spline models can be used, spline modeling is generally more convenient, faster, and leads to greater accuracies.

Examples with high-resolution data

Example 1 (Analyzing a high-resolution dataset simulated from a linear model). Here, we reconstruct a high-resolution but rather small dataset with 20000 data points (see Figure 2, left panel) generated from the OU model with drift function $\mu(x) = -\mu x$, diffusion function $\sigma(x) = \sigma$, parameter values $\mu = \sigma = 1$, and sampling time $\Delta = 0.01$. We obtain the parameter estimation $\hat{\mu} = -1.0586$ and $\hat{\sigma} = 0.9953$ for this linear parametric model. We also perform a spline reconstruction to this dataset using 8 equidistant knots (see Figure 2, right panel).

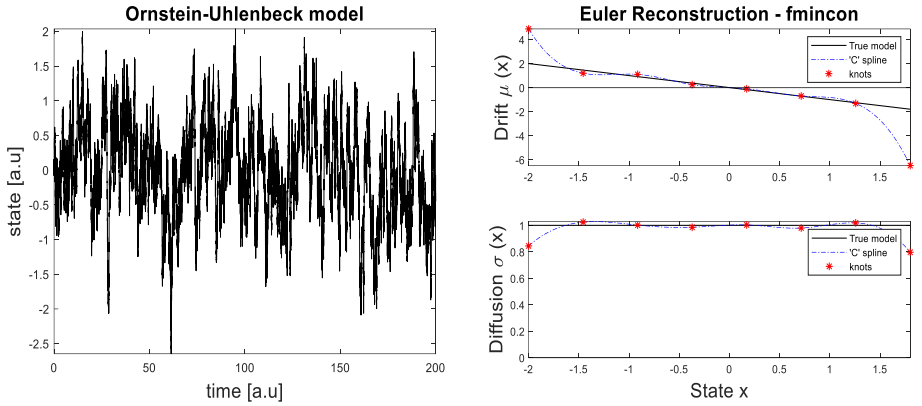

Figure 2. Illustration of spline reconstruction for a high-resolution dataset generated from a linear model. Data are illustrated in the left panel where 20000 data points, with sampling time $\Delta = 0.01$, are generated from the OU model. The right panel illustrates estimated cubic spline models (dot-dashed blue curves) for both the drift and diffusion functions using 8 regularly spaced knots over the state space alongside the true drift and diffusion functions (black curves). In figure captions for simulated data a.u. stands for arbitrary units.

Example 2 (Analyzing a small fraction of a high-resolution dataset). In this example, we have a very big dataset with 10^6 data points but we aim to use a small fraction of it which includes only 1% of the entire dataset, resulting in a smaller dataset with 10^4 data points. Similar to the first example the initial dataset is generated from the OU model with exactly the same model parameters and sampling time. In fact, the dataset in Example 1, corresponds to the first 20000 points of this larger dataset. Important to note is that both the original dataset and the smaller sample have the same resolution, so the sample dataset has a high-resolution, too. Reconstructing a small fraction of data is critical in situations where the original dataset is extremely big, making its full reconstruction time-consuming. However, a smaller, 'representative' fraction can still provide accurate results. For the smaller sample to be representative of the mother dataset, it should be a well-mixed random sample of what we call 'data pairs' which are consecutive pairs of data points (see red and blue open circles in Figure 3, left panel). To ensure the sample is a well-mixed representative we use stratified random sampling of data pairs, which is more randomized than simple random sampling (Cochran 1977). The results can vary slightly depending on the quality of the sample. For the sample illustrated in

Figure 3 (left panel), the parameter estimates are $\hat{\mu} = -1.035$ and $\hat{\sigma} = 0.997$. Figure 3 (right panel) depicts a spline reconstruction of this sample using 8 knots.

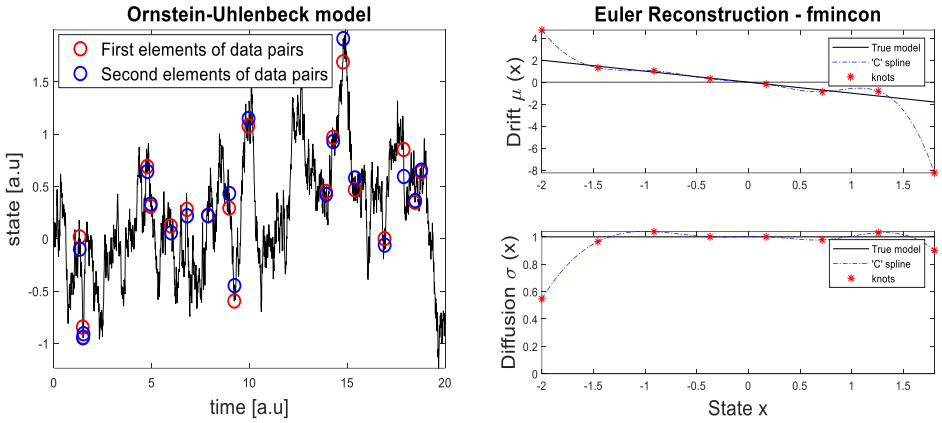

Figure 3. Illustration of spline reconstruction for a high-resolution random sample from a big dataset. For clarity, data are plotted up to time $T = 20$, although the simulation extends to time $T = 10^4$. The model parameters and data sampling time are identical to those in use in Figure 1. In the left panel, the open red and blue circles represent 'data pairs' which are consecutive data points selected based on a stratified random sampling from the mother dataset (the black time series). These data pairs form a small but well-mixed fraction, comprising only 1%, of the mother dataset. Right panel shows a spline reconstruction of this random sample, using 8 equidistant knots.

Example 3 (Analyzing a high-resolution dataset simulated from a nonlinear model). In this example we consider reconstructing a dataset (refer to Figure 4, left panel) simulated from the overgrazed model of May with drift function $\mu(x) = rx\left(1 - \frac{x}{K}\right) - \frac{\gamma x^2}{x^2 + a^2}$ and diffusion function $\sigma(x) = \sigma$ (in section 'Step4' the meaning of the parameters of May model were described). This dataset contains 10^5 data points with sampling time $\Delta = 0.01$. The parameter values are $r = 1.01, K = 10, \gamma = 2.75, a = 1.6, \sigma = 0.4$. Under these parameters the deterministic (i.e., without noise) model of May exhibits over-grazed and under-grazed alternative vegetation states (Figure 4, top right panel). Furthermore, the May model is nonlinear in terms of parameters K and a as well as the state variable x. Therefore, unlike the dataset in Example1, a longer dataset is needed here in order to be able to reconstruct transitions and time scales of shifts between

alternative basins of attraction. Nonetheless, we could estimate parameters with a rather good accuracy as $\hat{r} = 1.1805, \widehat{K} = 9.8416, \hat{\gamma} = 3.1242, \hat{a} = 1.5279, \hat{\sigma} = 0.3997$. Figure 4, right panel illustrates a cubic spline model fitted to this dataset.

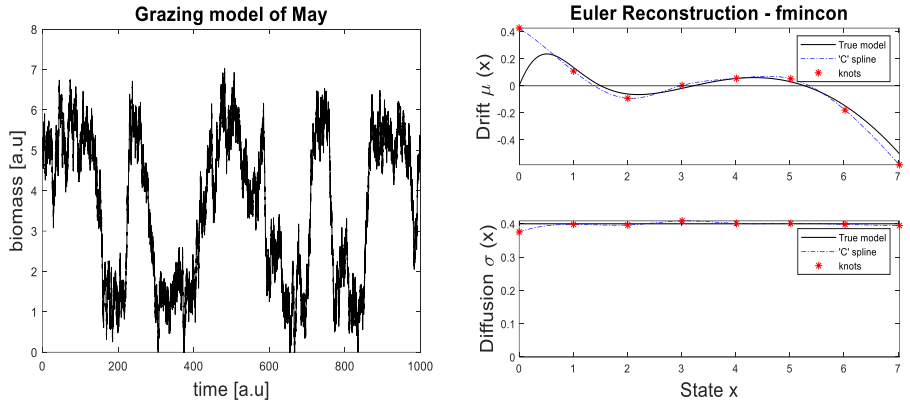

Figure 4. Illustration of spline modeling for a high-resolution dataset generated from a nonlinear model. Data are illustrated in the left panel where 10^5 data points, with sampling time $\Delta = 0.01$, are generated from the overgrazed model of May. The right panel illustrates estimated cubic spline models (dot-dashed blue curves) for both the drift and diffusion functions using 8 regularly spaced knots over the state space alongside the true drift and diffusion functions (black curves).

Example 4 (Analyzing a high-resolution ecological dataset). In this example we reconstruct a high-resolution ecological dataset. Data is a univariate index of Cyanobacteria biomass measured as phycocyanin concentrations in the Lake Mendota (Carpenter *et al.* 2020). The measurements were taken at minute intervals during the summer thermal stratification of 2011, a period known for common Cyanobacterial blooms (see Figure 5, left panel). For further details on this dataset, refer to references (Arani *et al.* 2021; Magnuson, Carpenter & Stanley 2023). This dataset does not meet one of the data requirements as it lacks Markov property (indicating high correlations at the measured time scale). However, a rarified sample of this dataset, consisting of every third data point (still maintaining high resolution), does exhibit Markovian behavior. Therefore, we applied Euler reconstruction using cubic splines to the rarified sample (see Figure 5, right panel). The usefulness of spline reconstruction becomes evident here: in real datasets

where choosing an appropriate model may be challenging, spline reconstruction proves to be a convenient solution.

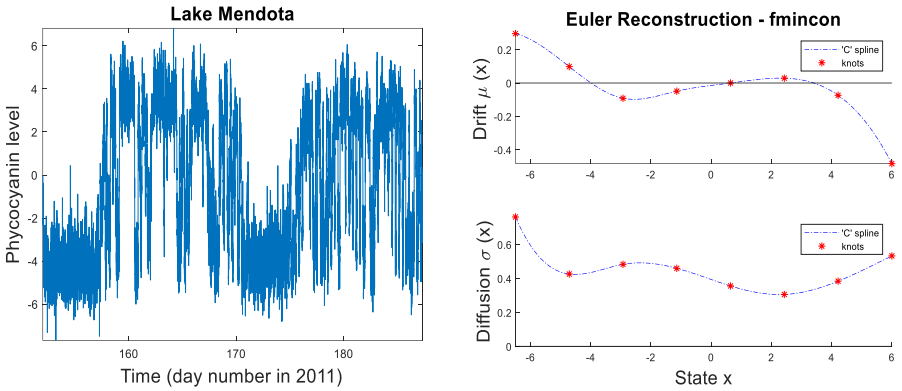

Figure 5. Illustration of spline modeling for a high-resolution ecological dataset. The left panel illustrates a high-resolution cyanobacterial measurement, taken at minute intervals, from lake Mendota. While this dataset does not meet one of the data requirements (i.e., Markov property), a rarified sample of this dataset, including every third data point, satisfies this requirement. In the right panel, cubic spline modeling was applied to this rarified sample to estimate the drift and diffusion functions. Data can be found online (Magnuson, Carpenter & Stanley 2023) with the URL https://doi.org/10.6073/pasta/fc8bd96677405945024ad708003be1fc

Examples with low-resolution data

Example 5 (Analyzing a low-resolution dataset with minimal resolution). Here, we reconstruct a dataset generated from the OU model using the same model parameters as in Example 1, i.e., $\mu = \sigma = 1$. This dataset is identical to the one in Example 1, but in that case, we only used the first 20000 data points. Although the original dataset contains 10^6 data points, we select every 100^{th} data points, resulting in a dataset with 100 times lower resolution and a total length of 10^4. The resulting low-resolution dataset has the lowest allowable resolution (with the sampling time of $\Delta = 1$) as its sampling time equals its relaxation time (see red dashed arrow in Figure 1). In practice it might still be possible to exceed this threshold resolution because we have only an approximate estimate of this critical limit of resolution. Nonetheless, as a proof of concept,

we show that it is possible to estimate the model parameters with high accuracy for this poor dataset by using Hermite reconstruction, a method specifically designed to tackle low-resolution data.

First, we applied a parametric reconstruction and estimated the model parameters with remarkable accuracy: $\hat{\mu} = -0.99683$ and $\hat{\sigma} = 0.99489$. Next, to implement the spline reconstruction, we used 'quadratic' splines (rather than the more common cubic splines) for the Hermite reconstruction, significantly reducing computational time and enhancing the likelihood of successful parameter estimation. This choice is rooted in the nature of Hermite reconstruction algorithm (as detailed in Appendices F, G, and particularly H). Once again, we could recover the true drift and diffusion functions with great accuracy. The Hermite reconstruction improved upon the outcomes of Euler reconstruction (Figure 6). For this extremely low-resolution dataset, we used 7 equidistant knots for both the drift and diffusion functions, compared to 8 knots used in other examples, to speed up the calculations. For Hermite reconstruction of very low-resolution data we need to be economical with respect to the number of knots as more knots means more model parameters. The major challenge for such 'highly damaged' datasets is to identify few candidate 'legitimate solutions' within the parameter space (see Step 6 in Methods section), to initiate optimization. Alternatively, one could consider using simpler additive models (i.e., models with constant diffusion function) with a larger number of knots for the drift function (see Example 7 or Example 9 in the tutorial).

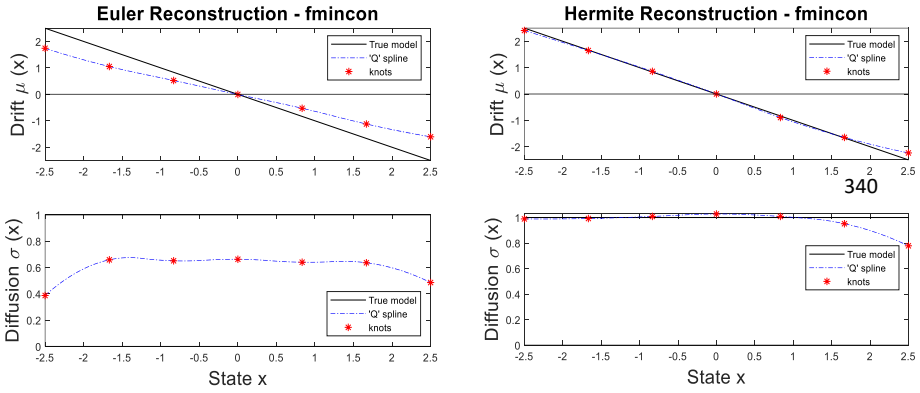

Figure 6. Illustration of spline modeling applied to an extremely low-resolution dataset generated from a linear model using Euler and Hermite reconstructions. The left panel illustrates Euler estimated quadratic spline models

(dot-dashed blue curves) for both the drift and diffusion functions, using 7 regularly spaced knots across the state space, alongside the true drift and diffusion functions (black curves). Similarly, the right panel illustrates Hermite estimated quadratic spline models (dot-dashed blue curves) for both the drift and diffusion functions using 7 regularly spaced knots across the state space, alongside the true drift and diffusion functions (black curves).

Example 6 (Analyzing a low-resolution dataset simulated from a nonlinear model). In this example, we reconstruct the same dataset as in Example 3, but with a resolution that is 300 times lower, resulting in a low-resolution dataset. To reconstruct it, we use Hermite reconstruction which offers higher accuracy than Euler reconstruction. For this, we apply a quadratic spline modeling technique. Despite the dataset's low-resolution, Hermite reconstruction yields accurate results in capturing the nonlinearities of the drift and diffusion functions (Figure 7, right panel), outperforming Euler reconstruction (Figure 7, left panel).

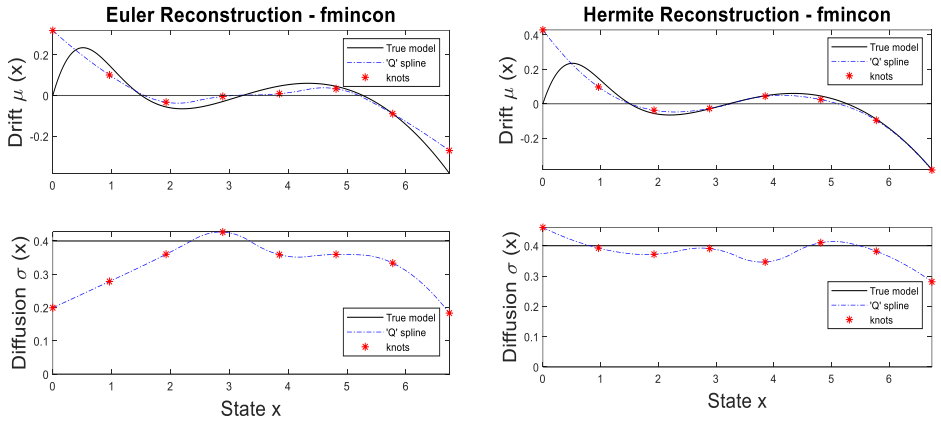

Figure 7. Illustration of Hermite and Euler reconstructions applied to a low-resolution dataset generated from a nonlinear model. The left panel illustrates Euler estimated quadratic spline models (dot-dashed blue curves) for both the drift and diffusion functions, using 8 regularly spaced knots across the state space, alongside the true drift and diffusion functions (black curves). The right panel illustrates Hermite estimated quadratic spline models (dot-dashed blue curves) for both the drift and diffusion functions, also using 8 regularly spaced knots over the state space, alongside the true drift and diffusion functions (black curves).

Example 7 (Analyzing a low-resolution and replicate dataset). In this example, we reconstruct a low-resolution dataset that differs from datasets analyzed so far. Instead of a single time series, we work with a fragmented dataset consisting of three fragments, referred to as 'replicate data'. Such data types are common in many practical applications, making it important to develop methods for their reconstruction as well. Here, we have three replicates (Figure 8, top panel) simulated from the grazing model of May. Each replicate is initialized at $x_0 = 8$ biomass and continues until perturbations drive it toward 0 biomass. In order to remove the impact of transient effects, the initial 5% of each replicate is discarded, and to further reduce the density of the data, every 30^{th} data point is selected.

The resulting replicate dataset not only has low resolution but also a very small total length of 471 data points (see solid dots in Figure 8, top panel). Reconstructing such data accurately is a significant challenge, and in such cases we aim to understand the qualitative, rather than quantitative, characteristics of the data-generating system. To address this, we fitted an additive (i.e., assuming the diffusion function is constant; See Figure 8, bottom panels) quadratic spline model to this poor-quality dataset. While Euler reconstruction produced inaccurate results (Figure 8, bottom left panel), Hermite reconstruction somewhat better (Figure 8, bottom right panel). Notably, Euler reconstruction failed to detect the presence of under-grazed states in the replicate dataset due to their low density in this dataset. In contrast, Hermite reconstruction was able to identify these states, at least qualitatively.

Example 8 (Analyzing a low-resolution and small climate dataset). In this example, a low-resolution climate dataset is reconstructed. The dataset, a $\delta^{18}O$ record from the North Greenland Ice Core Project (NGRIP) (2004), serves as a proxy for the temperature of the northern hemisphere (Kwasniok & Lohmann 2009), spanning the last 120 thousand years with a resolution of 20 years (Andersen *et al.* 2007). However, this dataset fails to meet two key data requirements outlined in Step2. Initially, the dataset exhibits non-stationarity, although it stabilizes within the period from 70 to 20 thousand years before the present (see Figure 9, top panel). During this epoch, the northern hemisphere climate witnessed alternating colder (stadial) and warmer (interstadial) states, attributed to Dansgaard–Oeschger (DO) events (Dansgaard *et al.*

1993). Within the specified time frame, the majority of DO events, from DO2 to DO18 out of a total of 25 DO events, occurred (2004). Secondly, the dataset lacks the Markov property, yet a rarified sample, with every other point demonstrates Markovian behavior approximately (refer to Table 1 in section 6.3 of the tutorial for further details).

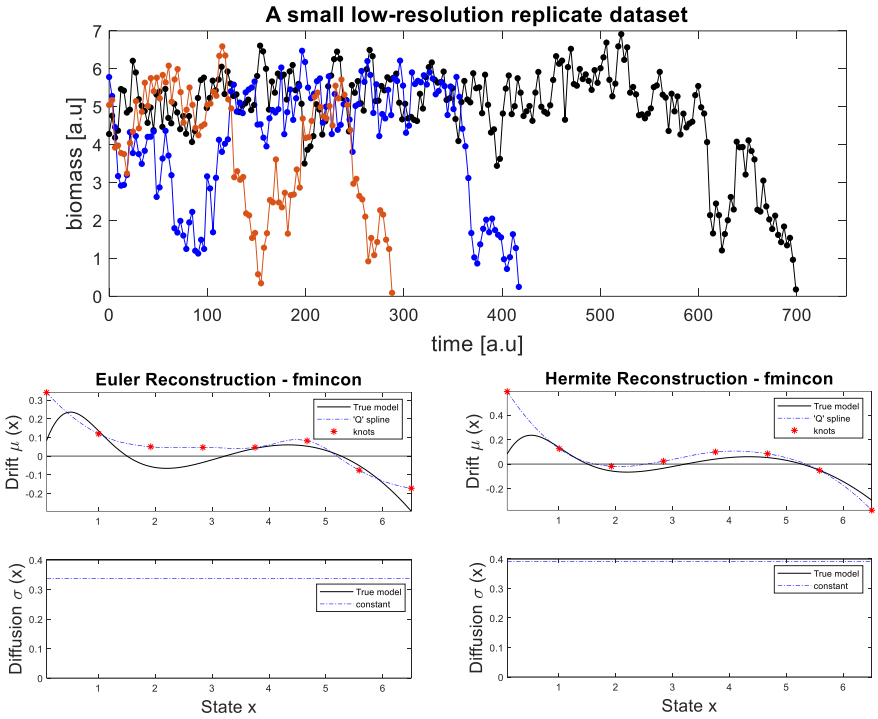

Figure 8. Illustration of Hermite and Euler reconstructions applied to a low-resolution and replicate dataset generated from a nonlinear model. The top panel illustrates three low-resolution replicate datasets. These replicates are generated using May's grazing model, where all replicate start at $x_0 = 8$ biomass and continue until perturbations drive them toward 0 biomass. To ensure the removal of transient effects, the first 5% of each replicate is discarded. Low-resolution data is then obtained by selecting every 30th data point from the replicates. The bottom left and bottom right panels are produced similarly to those in Figure 7.

Since this dataset has a low resolution, Hermite reconstruction is deemed suitable. Here, we employ quadratic spline modeling, using seven equidistant knots, to reconstruct the dataset. Figure 9 (bottom panels) illustrates the outcomes of Euler and Hermite reconstructions. Additionally, we introduce a significant and informative quantity known as '*effective potential*' (Arani *et al.* 2021) (For technical details refer to section 10.5 in the tutorial). Unlike deterministic systems where the location of equilibria can be identified using the drift function this is not the case with stochastic systems. For such systems effective potential should be used which is a quantity that incorporates information from both drift and diffusion functions. It is particularly useful for identifying alternative stable states, as is evident in this climate dataset. The minima of effective potential indicate the location of alternative stable states of stadial and interstadial states (solid dots in Figure 9, bottom right panel) which are separated by a repellor in between (open circle in Figure 9, bottom right panel).

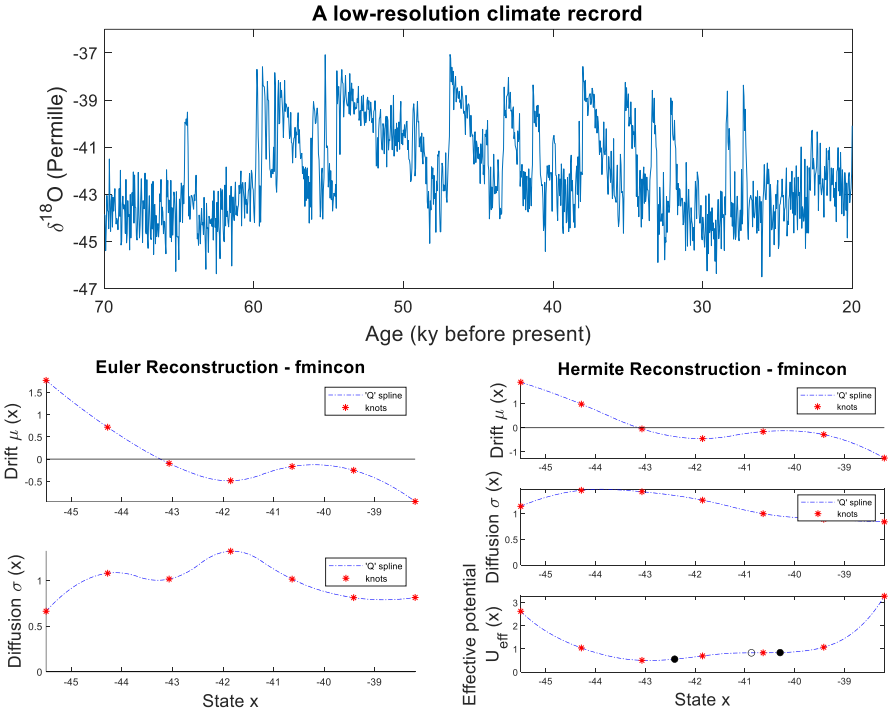

Figure 9. Illustration of Hermite and Euler reconstructions for a low-resolution climate dataset. The top panel illustrates a δ^{18}O climate record extending from 70 to 20 thousand years before the present time from NGRIP. This is used as a proxy for the temperature of the northern hemisphere which shows that the northern hemisohere climate alternated between cold stadial and warmer interstadial alternative climate states. In this time period a majority of Dansgaard-Oescher events occurred (see the labels 2 to 18). The description for bottom left and right panels are similar to that in Figure 6 or Figure 7. However, in the bottom right panel the effective potential is also depicted. Effective potential is useful to see whether there are alternative stable states in the dataset which is the case in this dataset (the solid dots represent alternative climate states of stadial and interstadial states separated by the open circle in between). We used seven equidistant knots to speed up the parameter estimation procedure.

Discussion

By delineating a systematic approach encompassing data preparation, model selection, and reconstruction techniques, we furnish researchers with a practical guide and MATLAB tools for analyzing diverse univariate datasets. The methods and examples presented in this paper furnish valuable insights into the reconstruction of Langevin systems from datasets of varying resolutions. Our discussion of data requirements is fundamental for reconstructing real datasets. We acknowledge the challenges posed by real datasets that do not meet these requirements and propose strategies such as data division (Example 9) or rarified sampling (Example 4 and Example 9) to address these limitations. We introduced two different modeling strategies: parametric versus spline modeling. The choice between parametric and spline modeling depends on the nature of the dataset and the researcher's familiarity with the underlying dynamics, including prior empirical knowledge. While parametric models offer interpretability with respect to model form, spline models provide flexibility in capturing unknown model nonlinearities, rendering them particularly suitable when the true model is uncertain. In general, we recommend to use spline modeling. Furthermore, without incorporation of splines, addressing Hermite reconstruction for low-resolution data pose considerable challenges.

The examples presented illustrate the application of our methodology to diverse datasets (e.g., high-resolution data, low-resolution data, Markov data, non-Markov data, big data, typical data, replicate data, etc.), spanning from simulated data to ecological data and climate records. These examples underscore the efficacy of both Euler and Hermite reconstruction techniques, demonstrating their utility across different resolutions and system complexities. Remarkably, Hermite reconstruction proves to be particularly valuable for low-resolution datasets, offering higher accuracy compared to Euler reconstruction. This is particularly important in the fields of ecology and climate sciences, as many ecological and climate datasets have low-resolution. We believe our 'MATLAB reconstruction package' together with a step-by-step and user-friendly tutorial is a valuable tool for ecologists and life scientists with little affinity for mathematical and statistical modeling.

Overall, our approach furnishes a systematic framework for reconstructing complex systems from observational data. While the examples provided demonstrate the efficacy of our methodology, further research is warranted to explore its applicability to other domains and datasets, especially those generated by more complex processes than Langevin models, such as diffusion-jump models (Gardiner 1985; Bandi & Nguyen 2003; Bandi & Phillips 2003) or models driven by Lévy noise (Siegert & Friedrich 2001; Li *et al*. 2022). Additionally, ongoing efforts to enhance computational efficiency and address computational challenges associated with multivariate Hermite reconstruction (Aït-Sahalia 2002) promise to advance the field further.

References

(2004) High-resolution record of Northern Hemisphere climate extending into the last interglacial period. *Nature,* **431,** 147-151.

Aït-Sahalia, Y. (2002) Closed-form likelihood expansions for multivariate diffusions. National Bureau of Economic Research Cambridge, Mass., USA.

Aït-Sahalia, Y.J.E. (2002) Maximum likelihood estimation of discretely sampled diffusions: a closed-form approximation approach. **70,** 223-262.

Andersen, K.K., Bigler, M., Buchardt, S.L., Clausen, H.B., Dahl-Jensen, D., Davies, S.M., Fischer, H., Goto-Azuma, K., Hansson, M.E. & Heinemeier, J. (2007) Greenland Ice Core Chronology 2005 (GICC05) and 20 year means of oxygen isotope data from ice core GRIP. *(No Title)*.

Anteneodo, C. & Queirós, S.D. (2010) Low-sampling-rate Kramers-Moyal coefficients. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics,* **82,** 041122.

Arani, B.M., Carpenter, S.R., Lahti, L., Van Nes, E.H. & Scheffer, M. (2021) Exit time as a measure of ecological resilience. *Science,* **372,** eaay4895.

Bakshi, G. & Ju, N.J.T.J.o.B. (2005) A Refinement to Aït-Sahalia's (2002)"Maximum Likelihood Estimation of Discretely Sampled Diffusions: A Closed-Form Approximation Approach". **78,** 2037-2052.

Bandi, F.M. & Nguyen, T.H. (2003) On the functional estimation of jump–diffusion models. *Journal of Econometrics,* **116,** 293-328.

Bandi, F.M. & Phillips, P.C. (2003) Fully nonparametric estimation of scalar diffusion models. *Econometrica,* **71,** 241-283.

Bolker, B.M., Gardner, B., Maunder, M., Berg, C.W., Brooks, M., Comita, L., Crone, E., Cubaynes, S., Davies, T. & de Valpine, P. (2013) Strategies for fitting nonlinear ecological models in R, AD M odel B uilder, and BUGS. *Methods in Ecology and Evolution,* **4,** 501-512.

Carpenter, S.R., Arani, B.M., Hanson, P.C., Scheffer, M., Stanley, E.H. & Van Nes, E. (2020) Stochastic dynamics of Cyanobacteria in long-term high-frequency observations of a eutrophic lake. *Limnology and Oceanography Letters,* **5,** 331-336.

Cochran, W.G. (1977) Sampling techniques. *Johan Wiley & Sons Inc*.

Connell, J.H. & Sousa, W.P. (1983) On the evidence needed to judge ecological stability or persistence. *The American Naturalist,* **121,** 789-824.

Dansgaard, W., Johnsen, S.J., Clausen, H.B., Dahl-Jensen, D., Gundestrup, N.S., Hammer, C.U., Hvidberg, C.S., Steffensen, J.P., Sveinbjörnsdottir, A. & Jouzel, J. (1993) Evidence for general instability of past climate from a 250-kyr ice-core record. *Nature,* **364,** 218-220.

De Boor, C. & De Boor, C. (1978) *A practical guide to splines*. springer-verlag New York.

Friedrich, R., Peinke, J., Sahimi, M. & Tabar, M.R.R.J.P.R. (2011) Approaching complexity by stochastic methods: From biological systems to turbulence. **506,** 87-162.

Gardiner, C.W. (1985) *Handbook of stochastic methods*. springer Berlin.

Hilborn, R. & Mangel, M. (2013) *The ecological detective: confronting models with data (MPB-28)*. Princeton University Press.

Honisch, C., Friedrich, R., Hörner, F. & Denz, C. (2012) Extended Kramers-Moyal analysis applied to optical trapping. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics,* **86,** 026702.

Kwasniok, F. & Lohmann, G. (2009) Deriving dynamical models from paleoclimatic records: Application to glacial millennial-scale climate variability. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics,* **80,** 066104.

Li, Y., Lu, Y., Xu, S. & Duan, J. (2022) Extracting stochastic dynamical systems with α-stable Lévy noise from data. *Journal of Statistical Mechanics: Theory and Experiment,* **2022,** 023405.

Magnuson, J.J., Carpenter, S.R. & Stanley, E.H. (2023) North Temperate Lakes LTER: High Frequency Data: Meteorological, Dissolved Oxygen, Chlorophyll, Phycocyanin-Lake Mendota Buoy 2006-current.

May, R.M. (1977) Thresholds and breakpoints in ecosystems with a multiplicity of stable states. *Nature,* **269,** 471-477.

Rinn, P., Lind, P.G., Wächter, M. & Peinke, J.J.a.p.a. (2016) The Langevin Approach: An R Package for Modeling Markov Processes.

Siegert, S. & Friedrich, R.J.P.R.E. (2001) Modeling of nonlinear Lévy processes by data analysis. **64,** 041107.

506