

# Boosting Bayesian Parameter Inference of Nonlinear Stochastic Differential Equation Models by Hamiltonian Scale Separation

Carlo Albert\*, Simone Ulzega\* and Ruedi Stoop†

May 4, 2018

## Abstract

Parameter inference is a fundamental problem in data-driven modeling. Given observed data that is believed to be a realization of some parameterized model, the aim is to find parameter values that are able to explain the observed data. In many situations, the dominant sources of uncertainty must be included into the model, for making reliable predictions. This naturally leads to stochastic models. Stochastic models render parameter inference much harder, as the aim then is to find a distribution of likely parameter values. In Bayesian statistics, which is a consistent framework for data-driven learning, this so-called posterior distribution can be used to make probabilistic predictions. We propose a novel, exact and very efficient approach for generating posterior parameter distributions, for stochastic differential equation models calibrated to measured time-series. The algorithm is inspired by re-interpreting the posterior distribution as a statistical mechanics partition function of an object akin to a polymer, where the measurements are mapped on heavier beads compared to those of the simulated data. To arrive at distribution samples, we employ a Hamiltonian Monte Carlo approach combined with a multiple time-scale integration. A separation of time scales naturally arises if either the number of measurement points or the number of simulation points becomes large. Furthermore, at least for 1D problems, we can decouple the harmonic modes between measurement points and solve the fastest part of their dynamics analytically. Our approach is applicable to a wide range of inference problems and is highly parallelizable.

## 1 Introduction

Modeling a dynamical process starts with a basic model that is usually obtained from a more or less deep insight into the nature of the process. The next step is the determination of the parameters of the model, based on observed data, which is generally a highly nontrivial task, in particular when complex behavior of such systems needs to be predicted, or when the measurements are noisy. A minimal example is a perceptron [22, 24], the basic element of a neuronal network, that predicts the double cosine value associated with the input of the corresponding sine function plus the sine's value at a fixed earlier time. While this task can easily be achieved for clean data using gradient descent learning, for noisy input data, this is largely impossible, as noise cannot be learned. The result is a distribution of potential parameter values (Fig. (1)).

---

\*Eawag, Swiss Federal Institute of Aquatic Science and Technology, 8600 Dübendorf, Switzerland.

†Institute of Neuroinformatics and Institute of Computational Science UZH/ETHZ, Irchel Campus 8057 Zurich, Switzerland.

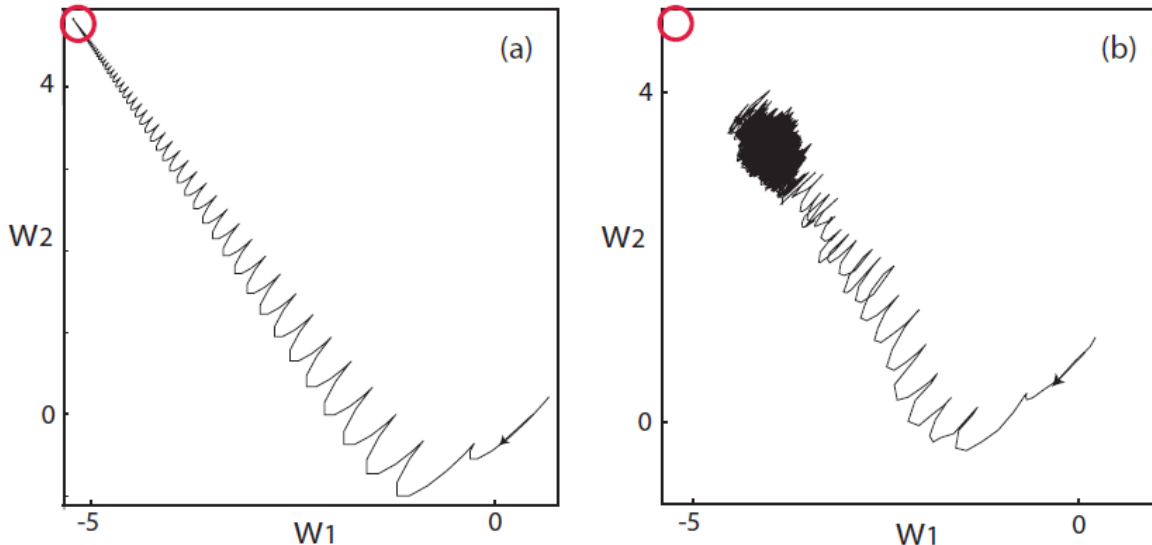


Figure 1: General problem setting: Parameter estimation (synaptic weights  $w_1, w_2$ ) of a perceptron, from a) noiseless, b) noisy data (noise sampled from a flat distribution over the interval  $[-0.2, 0.2]$ ). Whereas for noiseless data the estimates perfectly converge, for noisy data the estimates the system attempts to also include the noise, leading to a nontrivial distribution of the parameter estimates, and rendering the extraction of the optimal parameters a nontrivial task. Open circles: location of the optimal parameters in the noiseless case.

In Bayesian statistics, knowledge about parameters is expressed by probability distributions and learning is implemented as an update rule on these distributions (see, e.g. [3]). If a constant noise term is added to the output of a deterministic model, such as in the perceptron example above, Bayesian inference is straightforward. If noise enters the formulation of the model equations, however, Bayesian inference all of a sudden becomes computationally very expensive.

In our paper, we demonstrate the calibration of ordinary 1D stochastic differential equation (SDE) models based on noisy time series, and the quantification of the resulting parametric uncertainty. The generic approach that we use is exemplified by a simple SDE model from hydrology.

Problems of this kind are commonly solved by Monte Carlo (MC) methods that are based on simulating model realizations and comparing them to the data. Popular methods are particle filters [6, 15], Metropolis-within-Gibbs algorithms [28, 20] or Approximate Bayes Computations [17, 1, 30, 29]. A major problem with these simulation-based methods is, however, their inefficiency in the presence of many data points or high dimensions. One solution is to map the output space to a smaller dimensional space of summary statistics, and accept/reject proposed model parameters depending on how well associated model runs conform with the data in terms of these summary statistics [10]. However, how to choose the summary statistics to achieve a significant representation of the posterior parameter distribution is a largely unsolved problem.

These difficulties can be remedied with a reinterpretation of the Bayesian posterior distribution as the partition function of a statistical mechanics system and by simulating the dynamics of the latter. After discretizing the time of the original problem, we are led to a problem akin to the statistical mechanics of a polymer with harmonic bonds in an exterior potential [5]. In this framework, the measurements are interpreted as an

additional exterior potential that acts only on the polymer’s ‘measurement beads’ and confines their dynamics within the measurement uncertainty. The model parameters are interpreted as additional degrees of freedom coupling to all the beads of the polymer. To simulate the dynamics of this system, we apply the Hamiltonian Monte Carlo (HMC) algorithm [8], which combines Molecular Dynamics [2, 19] with the Metropolis algorithm [18]. Compared to traditional methods, the use of the Hamiltonian approach achieves much higher acceptance rates since data points are already used for the suggestion of new parameters, and thus model realizations incompatible with the data are never considered. The drawback is that the model equations need to be known and derivatives have to be calculated.

HMC requires two sets of parameters to be tuned: (i) the parameters that define the kinetic energy of the statistical mechanics system and (ii) the parameters that define the numerical integration scheme of Hamilton’s equation in the molecular dynamics part of the HMC algorithm. Efficiency of HMC algorithms can be gained if the kinetic term is made dependent on the configuration geometry of the statistical mechanics system. If the Riemann geometry of the parameter space of statistical models is taken into account, then the simulated search of paths across this manifold samples the target density in an utmost efficient way [12]. Unfortunately, this procedure is both demanding and computationally costly, depending strongly on the quality of the space’s extracted geometry.

Here we explore a computationally simpler approach, which, to our knowledge, has never been applied in the context of Bayesian inference before. Depending on the number of discretization points needed to approximate the original SDE system and the number of measurement points, the dynamics of the statistical mechanics system happen on very different time scales. This suggests a multiple time scale integration technique for the simulation of the statistical mechanics system [33]. We will show that for 1D SDE we can always find a parametrization, which decouples the harmonic modes in between measurement points from both the measurement points and the model parameters and allows for an analytical time-saving solution of the fastest part of the dynamics. Whilst for higher dimensional problems it is not always possible to solve part of the dynamics analytically, we believe that scale separation alone will render many SDE amenable to a full-fledged Bayesian inference with time-series. In fact, scale separation appears to be a generic feature if the dynamics of the SDE requires a large number of discretization points.

## 2 Inference Problem Setting

Consider, for simplicity and concreteness, a reservoir dynamics  $S(t)$  that on the observation time-scale is linear, with other (inflow and outflow) processes happening at much shorter time scales, so that they can be described by white noise. Furthermore, assume that this noise scales linearly with the system state  $S(t)$ . The model equation is thus given by the SDE

$$\dot{S}(t) = r(t) - \frac{1}{K} \left(1 + \frac{\gamma}{2}\right) S(t) + \sqrt{\frac{\gamma}{K}} S(t) \eta(t), \quad (1)$$

where  $r(t)$  denotes the time varying rain input,  $K$  denotes the retention time,  $\gamma$  is the noise strength, and  $\eta(t)$  indicates the white noise property, i.e.,

$$\langle \eta(t) \eta(t') \rangle = \delta(t - t'). \quad (2)$$

Eq. (1) is to be understood in the Stratonovich sense [25].

Properties of a transformed version of (1) have been derived, for constant input [9, 23, 11]. Here, suffice it to say that, for constant input  $r(t) = r_0$ , the equilibrium

distribution,  $P_{eq}(S)$ , is an inverse gamma distribution with scale parameter  $2Kr_0/\gamma$  and shape parameter  $(2 + \gamma)/\gamma$  (see Sect. 3.1), i.e.,

$$P_{eq}(S) \propto S^{-2(1+\gamma)/\gamma} e^{-2Kr_0/(\gamma S)}. \quad (3)$$

The mean of this expression equals the equilibrium solution of the unperturbed system ( $\gamma = 0$ )  $\langle S \rangle_{eq} = Kr_0$  and its variance, for  $\gamma < 2$ , is given by  $\langle (S - \langle S \rangle_{eq})^2 \rangle_{eq} = K^2 r_0^2 \gamma / (2 - \gamma)$ , which, for  $\gamma \geq 2$ , is seen to diverge. The power-law decay of the inverse gamma distribution is reminiscent of the invariance of Eq. (1) under re-scaling of both  $r(t)$  and  $S(t)$ . In real-world hydrology, for which Eq. (1) is a model, indeed often fat-tailed error distributions are observed [27].

While this equation was motivated by a popular hydrological model [4, 16], it is by no means restricted to this context. By means of the transformation  $S(t) = 1/n(t)$ , Eq. (1) turns into a model that has been suggested, e.g., as a phenomenological description of the dynamics of the neutron density in nuclear reactors [9].

In our setting, the input  $r(t)$  is a smooth and nowhere vanishing function. We assume the observed time-series,  $y_s$ , to be the outflow of the reservoir,  $S(t)/K$ , observed at times  $0 = t_1 < t_2 < \dots < t_{n+1} = T$ , with multiplicative independent log-normal errors,

$$\ln(y_s) = \ln\left(\frac{S(t_s)}{K}\right) + \sigma \epsilon_s, \quad s = 1, \dots, n+1, \quad (4)$$

where the  $\epsilon_s$  are uncorrelated standard normal errors. For simplicity, we assume  $\sigma$  as well as the input  $r(t)$  to be known, so that we are left with the task of inferring parameter combinations  $(K, \gamma)$  that are compatible with the data given by Eq. (4) in the Bayesian sense, where knowledge about parameters is expressed in terms of probability distributions.

We start our treatise by assuming that we have prior knowledge about a parameter vector,  $\boldsymbol{\theta}$ , in the form of a probability distribution,  $f_{\text{prior}}(\boldsymbol{\theta})$ , and measured data,  $\mathbf{y}$ , believed to be a realization of the model. The posterior knowledge, combining prior knowledge with the one acquired from data, is calculated by means of equation

$$f_{\text{post}}(\boldsymbol{\theta}|\mathbf{y}) = \frac{f_{\text{prior}}(\boldsymbol{\theta})L(\mathbf{y}|\boldsymbol{\theta})}{\int f_{\text{prior}}(\boldsymbol{\theta}')L(\mathbf{y}|\boldsymbol{\theta}')d\boldsymbol{\theta}'}, \quad (5)$$

where  $L(\mathbf{y}|\boldsymbol{\theta})$  is the probability distribution for model outputs given model parameters, evaluated at the measured data (the infamous *likelihood function*).

Before we set out to derive from Eqs. (1), (2) and (4) the likelihood function, we express the parameters and state variables by dimensionless quantities. Due to scale-invariance of the noise term,  $\gamma$  is already dimensionless. State variable  $S(t)$  and parameter  $K$  are replaced by the dimensionless quantities  $q(t)$  and  $\beta$ , respectively, which are defined by the transformations  $\beta = \sqrt{T\gamma/K}$  and  $S(t) = (T\gamma r(t)/\beta^2)e^{\beta q(t)}$ . In these new variables and parameters, Eq. (1) becomes the nonlinear SDE with constant noise

$$\dot{q}(t) = \frac{\beta}{T\gamma} e^{-\beta q(t)} - \frac{1}{T}\rho(t) + \frac{1}{\sqrt{T}}\eta(t), \quad (6)$$

with  $\rho(t) = (T/\beta)(d/dt)[\ln(r(t))] + (2 + \gamma)\beta/(2\gamma)$ .

The probability  $P(q_1, T|q_0, 0)$  of finding the system in a state  $q_1$  at time  $t = T$  if it was in an initial state  $q_0$  at time  $t = 0$ , is expressed as a *path-integral* as

$$P(q_1, T|q_0, 0) = \frac{1}{Z} \int e^{-S[q, \dot{q}]} \delta(q(T) - q_1) \delta(q(0) - q_0) \mathcal{D}q, \quad (7)$$

where the integral extends over all paths  $q : [0, T] \rightarrow \mathbb{R}$  and where the path-measure  $\mathcal{D}q$  is formally written as the infinite product  $\mathcal{D}q = \prod_t dq(t)$ . The *action* is a functional on

the space of paths and reads [14]

$$\mathcal{S}[q, \dot{q}] = \frac{1}{T} \int_0^T dt \left\{ \frac{1}{2} \left( T\dot{q}(t) + \rho(t) - \frac{\beta}{\gamma} e^{-\beta q(t)} \right)^2 - \frac{\beta^2}{2\gamma} e^{-\beta q(t)} \right\}. \quad (8)$$

This action includes the Jacobian that is introduced when changing coordinates from  $\eta(t)$  to  $q(t)$ .

If we denote the parameter vector  $\boldsymbol{\theta} = (\beta, \gamma)^T$  and assume a flat prior, the posterior (5) is, as a function of  $\boldsymbol{\theta}$ , proportional to the likelihood function

$$f_{\text{post}}(\boldsymbol{\theta}|\mathbf{y}) \propto \int \exp \left[ -\frac{1}{2} \sum_{s=1}^{n+1} \frac{(\ln(y_s/r(t_s)) - \beta q(t_s))^2}{\sigma^2} - \mathcal{S}[q, \dot{q}] \right] \mathcal{D}q. \quad (9)$$

Whereas the first term in the exponent describes the log-probability distribution of model outputs, for given model parameters, inputs and a system realization  $q(t_s)$ , the second term is the log-probability of the associated system realization  $q(t)$ .

When applying this approach now to real-world problems, instead of undertaking a prohibitive numerical computation of the path integral, we apply HMC to sample parameter vectors from a joint distribution of system realizations and model parameters given by an appropriate discretization of the action of the path-integral. By doing so, we observe that we obtain distinct regimes of time scales in the Hamiltonian that can be separated (see Sect. 3.2). This time scale separation simplifies and boosts our algorithm; in many cases it even permits parts of the required integrations to be done analytically.

## 3 Algorithm

### 3.1 Inference algorithm

For the inference algorithm, it is necessary to rewrite action (8) with the help of the time-dependent potential  $U(q, t) = \frac{1}{\gamma} e^{-\beta q} + q\rho(t)$  as

$$\begin{aligned} \mathcal{S}[q, \dot{q}] &= \frac{1}{T} \int_0^T dt \left\{ \frac{1}{2} T^2 \dot{q}^2(t) + \frac{1}{2} \left( \rho(t) - \frac{\beta}{\gamma} e^{-\beta q(t)} \right)^2 - T \frac{\partial U(q, t)}{\partial t} - \frac{\beta^2}{2\gamma} e^{-\beta q(t)} \right\} \\ &\quad + U(q(T), T) - U(q(0), 0) \\ &= \frac{1}{T} \int_0^T dt \left\{ \frac{1}{2} T^2 \dot{q}^2(t) + \frac{1}{2} \left( \rho(t) - \frac{\beta}{\gamma} e^{-\beta q(t)} \right)^2 - T q(t) \dot{\rho}(t) - \frac{\beta^2}{2\gamma} e^{-\beta q(t)} \right\} \\ &\quad + \frac{1}{\gamma} e^{-\beta q(T)} + q(T) \rho(T) - \frac{1}{\gamma} e^{-\beta q(0)} - q(0) \rho(0). \end{aligned} \quad (10)$$

With the action in this form, we easily derive the equilibrium distribution, for constant input  $r(t) = r_0$ , by plugging (7) and (10) into the detailed balance condition

$$P(q_1 t_1 | q_0 t_0) P_{eq}(q_0) = P(q_0 t_1 | q_1 t_0) P_{eq}(q_1), \quad (11)$$

and using the transformation  $q(t) \rightarrow q(-t)$ . We get, since  $\dot{\rho}(t) = 0$ ,

$$P_{eq}(q) \propto e^{-2U(q)}.$$

Back transformation to the original variables leads to Eq. (3).

For efficiently drawing parameter samples from (9), we interpret the latter as the partition function of a 1D statistical mechanics system and simulate its dynamics employing the HMC algorithm [8]. The model parameters  $\boldsymbol{\theta}$  are interpreted as additional dynamical

degrees of freedom coupling to the system variables  $q(t)$ . Each degree of freedom,  $q(t)$  and  $\boldsymbol{\theta}$ , is paired with a conjugate variable,  $p(t)$  and  $\boldsymbol{\pi}$  respectively, so that the system is defined by the Hamiltonian

$$\mathcal{H}_{\text{HMC}}(q, \boldsymbol{\theta}; p, \boldsymbol{\pi}) = K(p, \boldsymbol{\pi}) + V(q, \boldsymbol{\theta}), \quad (12)$$

where

$$K(p, \boldsymbol{\pi}) = \int_0^T \frac{p^2(t)}{2m(t)} dt + \sum_{\alpha=1}^2 \frac{\pi_{\alpha}^2}{2m_{\alpha}}, \quad (13)$$

and  $V(q, \boldsymbol{\theta})$  is the negative logarithm of the kernel of (9). The posterior (9) can then be expressed by the phase space path integral

$$f_{\text{post}}(\boldsymbol{\theta}|\mathbf{y}) \propto \int e^{-\mathcal{H}_{\text{HMC}}(q, \boldsymbol{\theta}; p, \boldsymbol{\pi})} \mathcal{D}p \mathcal{D}q d\boldsymbol{\pi}. \quad (14)$$

The HMC method, as a combination of the *Metropolis algorithm* [18] and *molecular dynamics* methods [2, 19], iterates the following steps:

1. Momenta  $p(t)$  and  $\boldsymbol{\pi}$  are sampled from the Gaussian distributions defined by Eq. (13).
2. The system is then allowed to evolve in  $(q, \boldsymbol{\theta}; p, \boldsymbol{\pi})$ -phase space for an arbitrary time interval  $\tau$  according to a volume-preserving and time-reversible solution of a discretized set of Hamilton equations.
3. The discretization error on the energy preservation due to the previous step is corrected by a Metropolis acceptance/rejection step.

The last step is the standard Metropolis algorithm, while the first two steps permit arbitrarily large jumps in phase space, while maintaining an arbitrarily large acceptance rate. Each new phase space configuration is associated with a combination of model parameters  $\boldsymbol{\theta}$ , which is compatible with the data in the Bayesian sense. Thus, omitting a possible burn-in, the parameter marginal of the simulated Markov chain of configurations represents a sample of the posterior probability distribution.

In order to simulate the dynamics of the Hamiltonian (12), we first need to discretize the path-integral (14). Let us assume that the measurement time points  $\{y_s\}_{s=1, \dots, n+1}$  of the time series (4) are equidistantly distributed on the time interval  $[0, T]$ , with  $t_1 = 0$  and  $t_{n+1} = T$ . Each interval between two consecutive data points is further partitioned into  $j$  bins, such that we have a total of  $nj + 1 = N \gg 1$  discretization points. The path-integral (14) is then approximated by an ordinary integral, with the approximate path-measure  $\mathcal{D}p \mathcal{D}q \approx \prod_i dp_i dq_i$ . The discretized versions of  $K(p, \boldsymbol{\pi})$  and  $V(q, \boldsymbol{\theta})$  are now given by

$$K(p, \boldsymbol{\pi}) \approx \sum_{i=1}^N \frac{p_i^2}{2m_i} \Delta t + \sum_{\alpha=1}^2 \frac{\pi_{\alpha}^2}{2m_{\alpha}}, \quad (15)$$

$$\begin{aligned} V(q, \boldsymbol{\theta}) \approx & \frac{\Delta t}{T} \sum_{i=2}^N \left\{ \frac{1}{2} T^2 \dot{q}_i^2 + \frac{1}{2} \left( \rho_i - \frac{\beta}{\gamma} e^{-\beta q_i} \right)^2 - \frac{\beta^2}{2\gamma} e^{-\beta q_i} - T q_i \dot{\rho}_i \right\} \\ & + \frac{1}{\gamma} e^{-\beta q_N} + q_N \rho_N - \frac{1}{\gamma} e^{-\beta q_1} - q_1 \rho_2 + \sum_{s=1}^{n+1} \frac{(\ln(y_s/r_{(s-1)j+1}) - \beta q_{(s-1)j+1})^2}{2\sigma^2}, \end{aligned} \quad (16)$$

with  $\dot{q}_i = (q_i - q_{i-1})/\Delta t$ ,  $\rho_i = T \ln(r(t_i)/r(t_{i-1})) / (\beta \Delta t) + (2 + \gamma)\beta / (2\gamma)$  and  $\dot{\rho}_i = (\rho_i - \rho_{i-1})/\Delta t$ , and where terms of order  $\mathcal{O}(N^{-1/2})$  were neglected. Note that we did not apply the mid-point discretization that is associated with the Stratonovich convention. In Eq.

(15) this leads to a different dynamics that does not, however, alter the posterior we are interested in, and in Eq. (16) we produce errors of the order  $\mathcal{O}(N^{-1/2})$  that we neglect.

Physically, the discretized Hamiltonian can be identified with a classical polymer chain of  $N$  beads with harmonic bonds between neighboring beads in an external field [5]. The latter consists of two parts, a field that results from the measurements and is felt by the measurement beads only (last term on the r.h.s. of Eq. (16)), and a field that results from the dynamics of the original Eq. (1) and is felt by all the beads. The masses  $m_i$  and  $m_\alpha$  are tunable parameters of the algorithm. Since measurement beads are constrained more than intermediate beads, we will assign larger masses to the former. Fig. (2) shows a typical realization of the dynamics of the polymer. Measurement beads only move within the measurement uncertainty, whilst the intermediate beads explore much larger regions of phase space.

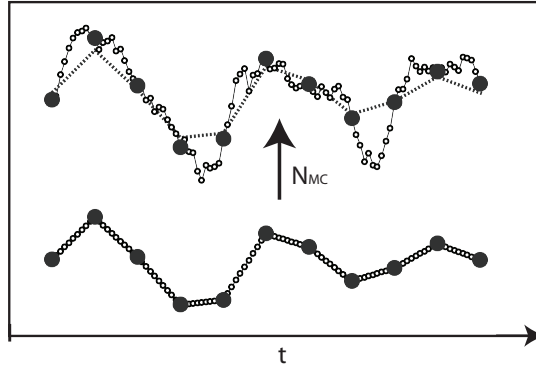


Figure 2: Simulated polymer chain dynamics, with  $n + 1 = 11$  data points (large circles) and  $j - 1 = 9$  intermediate beads (small circles). For other parameters see Sections 4 and 5. Bottom: initial state, where intermediate beads are on a linear interpolation between data points. Top: polymer after  $N_{MC} = 1000$  iterations of the propagation algorithm (dotted line: initial configuration). Clearly the new configuration is mostly determined by the dynamics of the light-mass intermediate beads, while the heavy-mass data points move to a much lesser extent.

We have thus reduced the original Bayesian inference problem to simulating the dynamics of a linear polymer (cf. Fig. (2)). Each state of this fictitious molecule corresponds to a well-defined configuration in the original phase space, characterized by a set of system variables  $\{q_i\}_{i=1,\dots,N}$  and a parameter vector  $\theta$ . It is now essential to note that potential (16) contains terms of distinct scaling in the potentially large numbers  $N$  and  $n$ , that refer to dynamics on distinct time-scales. In particular, for large  $N$ ,  $V(q, \theta)$  is dominated by its harmonic part, and to resolve its dynamics, brute force numerical integration of Hamilton's equations in step 2 of the HMC algorithm would require a very small discretization time-step.

### 3.2 Time scale separation

Whereas an interesting *approximate* approach would be to employ a partial averaging of the fast Fourier modes [7], we will use an *exact multiple time scale integration* based on Trotter's formula [33]. For this, we introduce so-called *staging variables*, and diagonalize the harmonic part between the measurement points. To this end, we rewrite the

discretized harmonic part of the action as

$$\sum_{i=2}^N \frac{T}{2\Delta t} (q_i - q_{i-1})^2 = \frac{T}{2} \sum_{s=1}^n \left\{ \frac{(q_{(s-1)j+1} - q_{sj+1})^2}{j\Delta t} + \sum_{k=2}^j \frac{k}{(k-1)\Delta t} (q_{(s-1)j+k} - q_{(s-1)j+k}^*)^2 \right\}, \quad (17)$$

with  $q_{(s-1)j+k}^* = ((k-1)q_{(s-1)j+k+1} + q_{(s-1)j+1})/k$ . The boundary beads, corresponding to the original measurement points, are not transformed,  $u_{sj+1} = q_{sj+1}$ ,  $s = 0, \dots, n$ , while, for the intermediate staging beads, we apply the coordinate transformations  $u_{sj+k} = q_{sj+k} - q_{sj+k}^*$ ,  $s = 0, \dots, n-1$ ,  $k = 2, \dots, j$ . Their inverse transformations are given by

$$q_{sj+1} = u_{sj+1}, \quad (18)$$

$$q_{sj+k} = \sum_{l=k}^{j+1} \frac{k-1}{l-1} u_{sj+l} + \frac{j-k+1}{j} u_{sj+1}, \quad (19)$$

which can be captured by the recursive relation

$$q_{sj+k} = u_{sj+k} + \frac{k-1}{k} q_{sj+k+1} + \frac{1}{k} u_{sj+1}. \quad (20)$$

The momenta are not transformed, which means we are using a non-canonical transformation. This alters only the dynamics of the system, not the posterior we are interested in.

We now split the Hamiltonian  $\mathcal{H}_{\text{HMC}}$  into components according to their scaling behavior in  $n$  and  $N$ , and write

$$\mathcal{H}_{\text{HMC}} = \mathcal{H}_N + \mathcal{H}_n + \mathcal{H}_1, \quad (21)$$

where

$$\mathcal{H}_N = \frac{1}{2} \sum_{s=1}^n \sum_{k=2}^j \left\{ \frac{\Delta t}{m'} p_{(s-1)j+k}^2 + \frac{Tk}{\Delta t(k-1)} u_{(s-1)j+k}^2 \right\}, \quad (22)$$

$$\begin{aligned} \mathcal{H}_n = & \frac{1}{2} \sum_{s=1}^{n+1} \left\{ \frac{\Delta t}{M} p_{(s-1)j+1}^2 + \frac{(\ln(y_s/r_{(s-1)j+1}) - \beta u_{(s-1)j+1})^2}{\sigma^2} \right\} \\ & + \frac{T}{2j\Delta t} \sum_{s=1}^n (u_{(s-1)j+1} - u_{sj+1})^2, \end{aligned} \quad (23)$$

$$\begin{aligned} \mathcal{H}_1 = & \sum_{\alpha=1}^2 \frac{\pi_\alpha^2}{2m_\alpha} + \frac{\Delta t}{T} \sum_{i=2}^N \left\{ \frac{1}{2} \left( \rho_i - \frac{\beta}{\gamma} e^{-\beta q_i} \right)^2 - \frac{\beta^2}{2\gamma} e^{-\beta q_i} - T q_i \dot{\rho}_i \right\} \\ & + \frac{1}{\gamma} e^{-\beta q_N} + q_N \rho_N - \frac{1}{\gamma} e^{-\beta q_1} - q_1 \rho_2. \end{aligned} \quad (24)$$

Here, we have introduced two masses,  $M$  and  $m'$ , for the boundary and staging beads, respectively. The scaling of these Hamiltonians can be derived from basic properties of discretized SDEs, from which we conclude that  $u_i \sim \sqrt{\Delta t}$ . Furthermore, in agreement with the *equipartition law* we find that  $p_i \sim 1/\sqrt{\Delta t}$ . Accordingly, we find that the harmonic part (22), for the staging beads, scales linearly with  $N$ . The terms of Eq. (23), including both the harmonic part for the boundary beads and the measurement term, scale linearly with  $n$ . Finally, Eq. (24) neither scales with  $n$  nor  $N$ . Thanks to the staging



variables,  $\mathcal{H}_N$  and  $\mathcal{H}_n$  have become fully decoupled. We use Trotter's formula [32] in order to design a reversible molecular dynamics integrator that takes the presence of the different time scales into account. For an appropriate partition of the Hamiltonian, three distinct regimes can be distinguished:

- (i)  $\mathcal{H}_N \sim \mathcal{H}_n \gg \mathcal{H}_1$ ,
- (ii)  $\mathcal{H}_N \gg \mathcal{H}_n \sim \mathcal{H}_1$ ,
- (iii)  $\mathcal{H}_N \gg \mathcal{H}_n \gg \mathcal{H}_1$ .

In the following we restrict ourselves to regime (ii), where the number of measurements  $n$  is assumed to be not too large and/or the measurement error  $\sigma$  to be not too small (the generalization of the method to the other regimes would, however, be straightforward). In this regime we may simply separate the harmonic part of the action for the staging beads from the rest and write

$$\mathcal{H}_{\text{HMC}} = \mathcal{H}_N + \mathcal{H}' . \quad (25)$$

For obtaining reversible integrators, we define the Liouville operators  $iL_N = \{\cdot, \mathcal{H}_N\}$ ,  $iL' = \{\cdot, \mathcal{H}'\}$ , where  $\{\cdot, \cdot\}$  denote the Poisson brackets that apply to functions on the phase space. Trotter's formula [31] allows us to write the Hamiltonian propagator as

$$e^{i(L_N + L')\tau} = (e^{iL_N(\Delta\tau/2)} e^{iL'\Delta\tau} e^{iL_N(\Delta\tau/2)})^P + \mathcal{O}(\tau^3/P^2), \quad (26)$$

for  $\tau = P\Delta\tau$ . Here, the outer propagator  $\exp[iL_N(\Delta\tau/2)]$  reflects much faster dynamics than the inner one. However, thanks to our re-parametrization, it describes the dynamics of uncoupled harmonic oscillators, which we can readily solve. Masses and frequencies of the oscillators are derived from (22) as

$$m = m'/\Delta t, \quad \omega_k = \sqrt{\frac{Nk}{(k-1)m}} . \quad (27)$$

The outer propagator becomes

$$u_{(s-1)j+k}(\Delta\tau/2) = u_{(s-1)j+k}(0) \cos(\omega_k \Delta\tau/2) + \frac{p_{(s-1)j+k}(0)}{m\omega_k} \sin(\omega_k \Delta\tau/2), \quad (28)$$

$$p_{(s-1)j+k}(\Delta\tau/2) = p_{(s-1)j+k}(0) \cos(\omega_k \Delta\tau/2) - m\omega_k u_{(s-1)j+k}(0) \sin(\omega_k \Delta\tau/2), \quad (29)$$

for  $s = 1, \dots, n$  and  $k = 2, \dots, j$ . For the inner propagator, we employ the time-reversible and volume preserving velocity Verlet algorithm [26], which leads for the boundary beads to

$$u_{(s-1)j+1}(\Delta\tau) = u_{(s-1)j+1}(0) + \frac{\Delta\tau}{M} p_{(s-1)j+1}(0) + \frac{\Delta\tau^2}{2M} F_{(s-1)j+1}[\mathbf{u}(0), \boldsymbol{\theta}(0)], \quad (30)$$

$$p_{(s-1)j+1}(\Delta\tau) = p_{(s-1)j+1}(0) + \frac{\Delta\tau}{2} (F_{(s-1)j+1}[\mathbf{u}(0), \boldsymbol{\theta}(0)] + F_{(s-1)j+1}[\mathbf{u}(\Delta\tau), \boldsymbol{\theta}(\Delta\tau)]), \quad (31)$$

with  $s = 1, \dots, n+1$  and where  $F_i[\mathbf{u}, \boldsymbol{\theta}]$  denotes the partial derivative of  $\mathcal{H}'[\mathbf{u}, \boldsymbol{\theta}]$  w.r.t.  $u_i$ . Analogous equations emerge for the model parameters  $\boldsymbol{\theta}$  and their momenta  $\boldsymbol{\pi}$ , by exchanging  $u$  by  $\theta$  and  $p$  by  $\pi$ , along with the corresponding masses, respectively. For the staging beads only the momenta need to be updated (because the associated kinetic term is not part of  $\mathcal{H}'$ , but of  $\mathcal{H}_N$ ). Thus, with  $s = 1, \dots, n$  and  $k = 2, \dots, j$  as before,

$$p_{(s-1)j+k}(\Delta\tau) = p_{(s-1)j+k}(0) + \frac{\Delta\tau}{2} (F_{(s-1)j+k}[\mathbf{u}(0), \boldsymbol{\theta}(0)] + F_{(s-1)j+k}[\mathbf{u}(\Delta\tau), \boldsymbol{\theta}(\Delta\tau)]). \quad (32)$$

The propagators (28) through (32) are applied sequentially  $P$  times to calculate the system evolution over time  $\tau$ . The proposed configuration,  $(\mathbf{u}', \boldsymbol{\theta}'; \mathbf{p}', \boldsymbol{\pi}')$  is accepted with Metropolis probability  $\min\left(1, e^{\mathcal{H}_{\text{HMC}}(\mathbf{u}, \boldsymbol{\theta}; \mathbf{p}, \boldsymbol{\pi}) - \mathcal{H}_{\text{HMC}}(\mathbf{u}', \boldsymbol{\theta}'; \mathbf{p}', \boldsymbol{\pi}')}\right)$ . The next iteration then starts with sampling a new momentum vector  $(\mathbf{p}, \boldsymbol{\pi})$ .

The analytical solution (28) and (29) is one main boosting part of our algorithm. To find such a solution, it is important to arrive at model equations of the form (6), where the noise term neither depends on the state variables, nor on the parameters to be inferred. In a one dimensional model this can always be achieved through re-parametrization (see, e.g., chapter 5 in [21]). In higher dimensions, this will not always be possible. But even in such cases, we will be able to boost our algorithm through assigning smaller time intervals  $\Delta\tau$  to the fast dynamics and larger ones to the slow dynamics.

## 4 Results

For our toy system, we have considered a simple sinusoidal input  $r(t) = \sin^2(0.01t) + 0.1$ . A system realization was first obtained from Eq. (1) using  $K_{\text{true}} = 50$  (in arbitrary units of time) and  $\gamma_{\text{true}} = 0.2$ . Such system realization was then used to generate a synthetic time series of observed data according to Eq. (4). The error  $\sigma$  was set to 0.1. The input signal, the "true" system realization and the corresponding data time series are shown in Fig. (3). A set of 200 system realizations sampled from the integrand of Eq. (9), based on

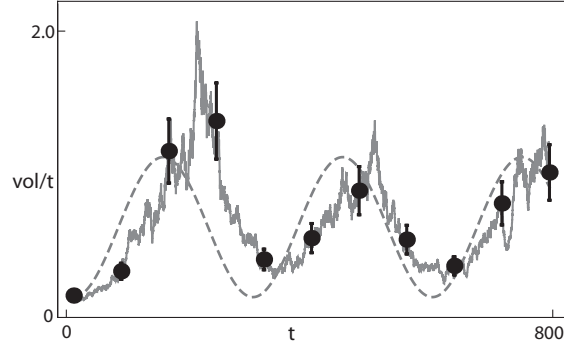


Figure 3: System realization (solid line) with synthetic observations (filled circles, with error bars indicating the assumed measurement uncertainty). The system response closely follows the oscillations of the sinusoidal input (dashed) in a time-delayed manner. Parameters for this figure and all following figures:  $K_{\text{true}} = 50$  and  $\gamma_{\text{true}} = 0.2$ .

$n + 1 = 11$  measurement points and  $N = 301$  discretization points, is shown in Fig. (4), together with the generated synthetic data. These samples were generated with a HMC algorithm, where the masses were set (in arbitrary units) to  $M = 720$  for the measurement beads, to  $m' = 130$  for the intermediate discretization beads, and to  $m_\alpha = 150$  for both the dimensionless parameters  $\beta$  and  $\gamma$ . The different dynamics of the heavy measurement beads and the light discretization beads can be appreciated in Fig. (4).

The Markov chains for parameters  $K$  and  $\gamma$ , obtained after  $N_{MC} = 50000$  iterations of the HMC algorithm, are shown in Figs. (5) and (6), respectively.

The efficiency of the algorithm can be appreciated best by inspecting the system evolution in the phase space  $K - \gamma$  (Fig. (7)). The starting point of the algorithm was set to a linear interpolation of the data points, together with values for the model parameters that were deliberately chosen far off the truth (open circle in Fig. 7). Nevertheless, the very first step of the algorithm already takes the system to the vicinity of the true

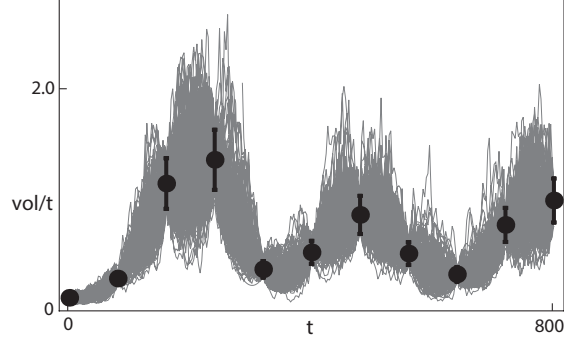


Figure 4: Simulated system realizations associated with synthetic data, based on  $n + 1 = 11$  measurement points and  $N = 301$  discretization points.

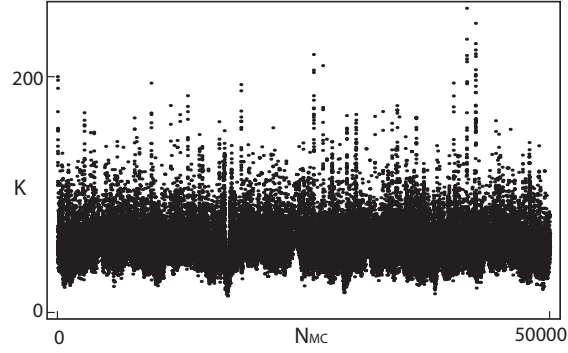


Figure 5: Markov chain evolution of the inferred parameter  $K$ .

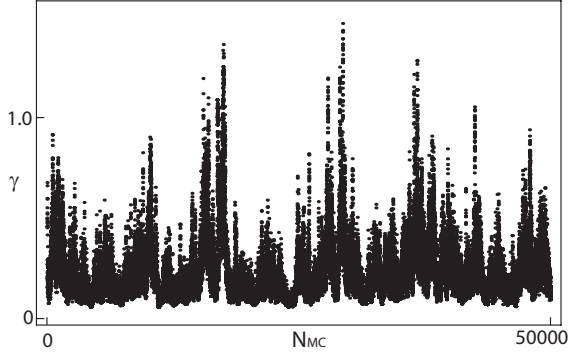


Figure 6: Markov chain evolution of the inferred parameter  $\gamma$ .

parameter values, where most of its dynamics then occurs. Few excursions lead far away from the true parameter values. These explore the heavy tails of the posterior parameter distribution.

The Markov chains (Figs. (5) and (6)) determine the probability density functions (PDF) for  $K$  and  $\gamma$ , respectively. The results obtained by using the built-in kernel density estimator provided by Mathematica (version 10) are exhibited in Fig. (8); they are fully compatible with the true, to be inferred, parameter values  $K_{\text{true}}$  and  $\gamma_{\text{true}}$ .

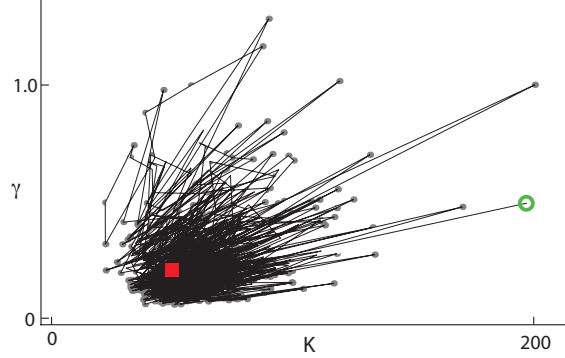


Figure 7: System dynamics in the phase space  $K - \gamma$ . The circle represents the initial state, while the square corresponds to the true parameter values used to generate the data.

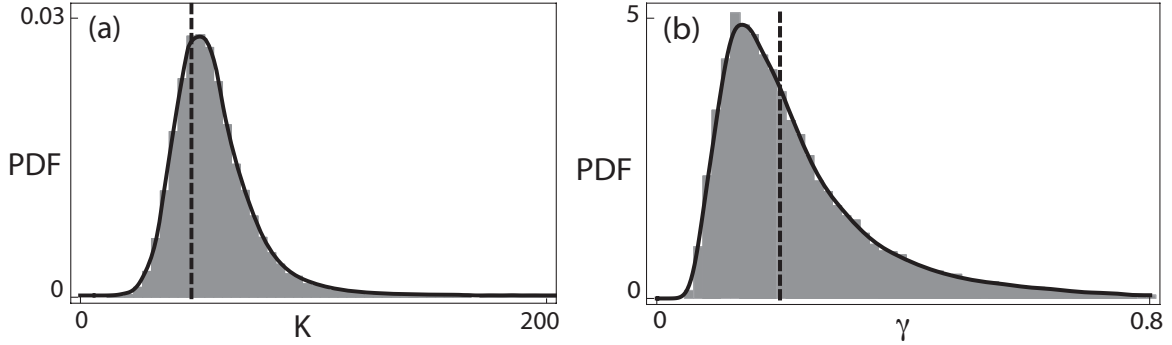


Figure 8: Probability density functions for the inferred parameters, (a)  $K$  and (b)  $\gamma$ . The true values used to generate the data are represented by the dashed vertical lines.

## 5 Implementation

The algorithm was implemented in C++ (version C++11) using the open source *Adept* library (version 1.1; [13]), which provides a powerful tool for fast reverse-mode automated differentiation (AD). Our algorithm benefits greatly from the use of AD. This gives us the possibility to modify Eq. (1) and therefore the action (10), while leaving the implementation of the algorithm unaltered. This makes our program extremely flexible and suitable for a much broader range of applications than the simple exemplary SDE model described here.

The simulations were run on both serial and parallel implementations of the algorithm on a 64-bit Linux system equipped with two 12-core 2.7 GHz processors (Intel Xeon E5-2697v2) and 64 GB of memory clocked at 1866 MHz. We used  $n + 1 = 11$  measurement points and  $N = 101, 201, \dots, 501$  discretization points. In the Hamiltonian propagator (26), we set  $\Delta\tau = 0.25$  and  $P = 3$ , with a constant total observation time  $T = 833$  (arbitrary units of time). The initial values of the parameters were set to  $K = 200$  and  $\gamma = 0.5$ .

For example, a complete run with  $N_{MC} = 50000$  iterations with  $n = 10$  and  $N = 301$  required about 43 seconds with the serial implementation of the algorithm. In the case of our toy system the burn-in phase is extremely short and can be safely ignored. Under these conditions the algorithm can be parallelized in a straightforward manner simply by breaking up the Markov chain into several smaller independent chains. The execution time

with a fixed-size problem scales in a reasonably linear way with the number of processes (strong scaling). Our example could be therefore run in only about 3 seconds using 16 processors. An alternative strategy, suitable for long time-series, would be to parallelize the updates of the polymer beads in each step of a single MC chain.

## 6 Conclusions

We presented a novel, extremely efficient and versatile approach for data-based SDE parameter estimation. Our algorithm obtains its strength from translating the problem of generating posterior parameter samples into the problem of simulating the dynamics of a statistical mechanics system; the main novelty in our algorithm is the exploitation of the fact that this dynamics generically happens on very different time scales. Furthermore, at least for 1D systems, our approach also allows for an analytical, and therefore computationally efficient, integration of the fastest part of the dynamics.

In most application cases, our choice of a fixed diagonal mass matrix, for reasonable choices of masses - heavier for the measurement beads, lighter for the discretization beads - can be expected to work well. Nonetheless, if the curvature of the potential varies strongly, it might be beneficial to adapt the mass matrix to the local curvature as suggested in [12]. For such cases, a combination of the scale separation method proposed in this paper with the local mass matrix adaptation of [12] might be the most efficient solution. This extension, however, comes at the price of a computational overhead (second derivatives have to be calculated and implicit equations have to be solved), and we will no longer be able to solve part of the dynamics analytically. On a more general level, given the wide field of very distinct applications, optimal parameter inference for SDE models will not be provided by one single approach, but will require a set of tools, to make the optimal choice from. We expect statistical physics to continue making strong contributions toward this aim.

The structure of our algorithm is well suited to parallelization, which is important in particular if we deal with a high number of measurements. The algorithm can easily be adapted to other inference problems, such as higher dimensional SDE, and SDE coupled to ODE. These adaptations and extensions will, however, be addressed in future works.

## Funding

This work was partly financed by the Eawag Discretionary Fund.

## References

- [1] C. Albert, H. R. Künsch, and A. Scheidegger. A Simulated Annealing Approach to Approximate Bayes Computations. *Stat. Comput.*, 25(6):1217–1232, 2015.
- [2] B. J. Alder and T. E. Wainwright. Studies in Molecular Dynamics. I. General Method. *J. Chem. Phys.*, 31:459–466, 1959.
- [3] G. EP. Box and G. C. Tiao. *Bayesian inference in statistical analysis*, volume 40. John Wiley & Sons, 2011.
- [4] A. Breinholt, F.O. Thordarson, J.K. Moller, M. Grum, P.S. Mikkelsen, and H. Madsen. Grey-box modelling of flow in sewer systems with state-dependent diffusion. *Environmetrics*, 22(8):946–961, 2011.
- [5] D. Chandler and P. G. Wolynes. Exploiting the isomorphism between quantum theory and classical statistical mechanics of polyatomic fluids. *J. Chem. Phys.*, 74(7):4078–4095, 1981.

- [6] N. Chopin, P. E. Jacob, and O. Papaspiliopoulos. SMC2: an efficient algorithm for sequential analysis of state space models. *J. Roy. Stat. Soc. B*, 75(3):397–426, 2013.
- [7] JD Doll, Rob D Coalson, and David L Freeman. Fourier path-integral Monte Carlo methods: Partial averaging. *Physical review letters*, 55(1):1, 1985.
- [8] S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Phys. Lett. B*, 195(2):216–222, 1987.
- [9] WL Dutré and AF Debosscher. Exact Statistical Analysis of Nonlinear Dynamic Nuclear-Power Reactor Models by the Fokker-Planck Method Part I: Reactor with Direct Power Feedback. *Nuclear Science and Engineering*, 62(3):355–363, 1977.
- [10] P. Fearnhead and D. Prangle. Constructing summary statistics for approximate Bayesian computation: semi-automatic approximate Bayesian computation. *J. Roy. Stat. Soc. B*, 74(3):419–474, 2012.
- [11] H. Fujisaka, H. Ishii, M. Inoue, and T. Yamada. Intermittency caused by chaotic modulation. iilyapunov exponent, fractal structure and power spectrum. *Progress of theoretical physics*, 76(6):1198–1209, 1986.
- [12] M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. Roy. Stat. Soc. B*, 73(Part 2):123–214, 2011.
- [13] R. J. Hogan. Fast reverse-mode automatic differentiation using expression templates in C++. *ACM Trans. Math. Softw.*, 40(4):1–16, 2014.
- [14] A. W. C. Lau and T. C. Lubensky. State-dependent diffusion: thermodynamic consistency and its path integral formulation. *Phys. Rev. E*, 76(1):011123, 2007.
- [15] X. Liu and M. Niranjana. State and parameter estimation of the heat shock response system using Kalman and particle filters. *Bioinformatics*, 28(11):1501–1507, 2012.
- [16] V. Livina, Y. Ashkenazy, Z. Kizner, V. Strygin, A. Bunde, and S. Havlin. A stochastic model of river discharge fluctuations. *Physica A: Statistical Mechanics and its Applications*, 330(1):283–290, 2003.
- [17] J.M. Marin, P. Pudlo, C.P. Robert, and R.J. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6, SI):1167–1180, 2012.
- [18] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equation of state calculations by fast computing machines. *J. Chem. Phys.*, 21(6):1087–1092, 1953.
- [19] A. Rahman. Correlations in the Motion of Atoms in Liquid Argon. *Physical Review*, 136:405–411, 1964.
- [20] P. Reichert and J. Mieleitner. Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters. *Water Resources Res.*, 45, 2009.
- [21] H. Risken. *The Fokker-Planck Equation; Methods of Solution and Applications; 2nd ed.* Springer, 1989.
- [22] F. Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.
- [23] A. Schenzle and H. Brand. Multiplicative stochastic processes in statistical physics. *Phys. Rev. A*, 20(4):1628, 1979.
- [24] W.H. Steeb, Y. Hardy, A. Hardy, and R. Stoop. *Problems & Solutions In Scientific Computing With C++ And Java Simulations*. World Scientific Publishing Co., Inc., 2004.
- [25] R. L. Stratonovich. *Conditional Markov Processes and their Application to the Theory of Optimal Control*. Elsevier, New York, 1968.

- [26] W. C. Swope, H. C. Andersen, P. H. Berens, and K. R. Wilson. A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters. *J. Chem. Phys.*, 76(1):637–649, 1982.
- [27] M. Thyer, B. Renard, D. Kavetski, G. Kuczera, S. W. Franks, and S. Srikanthan. Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. *Water Resources Res.*, 45(12), 2009.
- [28] L. Tomassini, P. Reichert, H. R. Künsch, C. Buser, R. Knutti, and M. E. Borsuk. A smoothing algorithm for estimating stochastic, continuous time model parameters and its application to a simple climate model. *J. Roy. Stat. Soc. C*, 58(5):679–704, 2009.
- [29] T. Toni and M.P.H. Stumpf. Simulation-based model selection for dynamical systems in systems and population biology. *Bioinformatics*, 26(1):104–110, 2010.
- [30] T. Toni, D. Welch, N. Strelkowa, A. Ipsen, and M. P. H. Stumpf. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface*, 6(31):187–202, 2009.
- [31] H. F. Trotter. On the product of semi-groups of operators. *Proc. Amer. Math. Soc.*, 10(4):545–551, 1959.
- [32] M. E. Tuckerman, B. J. Berne, and G. J. Martyna. Reversible multiple time scale molecular dynamics. *J. Chem. Phys.*, 97(3):1990–2001, 1992.
- [33] M. E. Tuckerman, B. J. Berne, G. J. Martyna, and M. L. Klein. Efficient molecular dynamics and hybrid Monte Carlo algorithms for path integrals. *J. Chem. Phys.*, 99(4):2796–2808, 1993.