Systematic Physics Constrained Parameter Estimation of Stochastic Differential Equations

Daniel Peavoy[a,1], Christian L. E. Franzke[b,*], Gareth O. Roberts[c]

[a] *Complexity Science Centre, University of Warwick, Coventry, UK*
[b] *Meteorological Institute and Center for Earth System Research and Sustainability (CEN), University of Hamburg, Hamburg, Germany*
[c] *Department of Statistics, University of Warwick, Coventry, UK*

Abstract

A systematic Bayesian framework is developed for physics constrained parameter inference of stochastic differential equations (SDE) from partial observations. Physical constraints are derived for stochastic climate models but are applicable for many fluid systems. A condition is derived for global stability of stochastic climate models based on energy conservation. Stochastic climate models are globally stable when a quadratic form, which is related to the cubic nonlinear operator, is negative definite. A new algorithm for the efficient sampling of such negative definite matrices is developed and also for imputing unobserved data which improve the accuracy of the parameter estimates. The performance of this framework is evaluated on two conceptual climate models.

Keywords: Global Stability, Parameter Inference, Imputing Data, Stochastic Differential Equations, Physical Constraints, Stochastic Climate Models
2010 MSC: 86-08,
2010 MSC: 65C60,
2010 MSC: 86A10

1. Introduction

In many areas of science the inference of reduced order hybrid dynamic-stochastic models, which take the form of stochastic differential equations (SDE), from data is very important. For many applications running full resolution dynamical models is computationally prohibitive and in many situations one is mainly interested in large-scale features and not the exact evolution of the fast, small scale features, which typically determine the time step size. Thus, reduced order stochastic models are an attractive alternative. Examples are molecular dynamics (21), engineering turbulence (20) and climate science (14, 15, 22).

The inference of such models has been done using non-parametric methods (40, 17, 8) from partial observations. These non-parametric methods need very long time series for reliable parameter estimates and can be used only for very low-dimensional models because of the curse

[*] Center for Earth System Research and Sustainability (CEN), University of Hamburg, Grindelberg 7, 20144 Hamburg, Germany, christian.franzke@gmail.com

[1] Present affiliation: Manufacturing Technology Centre, Coventry, UK

of dimension. More importantly, they do not necessarily obey conservation laws or stability properties of the full dimensional dynamical system. In many areas of science one can derive reduced order models from first principles (26, 31) such that certain fundamental properties of the full dynamics are still valid. These methods provide us with parametric forms for the model fitting. Physical constraints then not only constrain the parameters one has to estimate but they can also ensure global stability. Thus, there is a need for systematic physics constrained model and parameter estimation procedures (32, 28).

For instance, the climate system is governed by conservation laws like energy conservation. Based on this energy conservation property the normal form of stochastic climate models has been derived by (26) using the stochastic mode reduction procedure (23, 24, 25, 14, 15). This procedure allows the systematic derivation of reduced order models from first principles. This normal form provides a parametric form for parameter estimation from partial observations which we will use in this study.

The fundamental form of climate models is given by

$$\frac{d\mathbf{z}}{dt} = F + L\mathbf{z} + B(\mathbf{z}, \mathbf{z}), \tag{1}$$

where $\mathbf{z} \in \mathbb{R}^N$ denotes the N-dimensional state vector, F the external forcing, L a linear and B a quadratic nonlinear operator. The nonlinear operator B is conserving energy $\mathbf{z} \cdot B(\mathbf{z}, \mathbf{z})$. For current climate models N is of the order of $10^6 - 10^8$. This shows that running complex climate models is computationally expensive. But for many applications like extended-range (periods of more than 2 weeks), seasonal and decadal climate predictions one is only interested in the large-scale circulation of the climate system and not whether there will be a cyclone over London on a particular day next year. The large-scale circulation can successfully be predicted using reduced order models (37, 1, 14, 15, 22).

The stochastic mode reduction procedure (23, 24, 25) provides a systematic framework for deriving reduced order climate models with a closure which takes account of the impact of the unresolved modes on the resolved modes. In order to derive reduced order models one splits the state vector $\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$ into resolved $\mathbf{x}$ and unresolved $\mathbf{y}$ modes. The stochastic mode reduction procedure now enables us to systematically derive a reduced order climate model which only depends on $\mathbf{x}$

$$d\mathbf{x} = \left(\tilde{F} + \tilde{L}\mathbf{x} + \tilde{B}(\mathbf{x}, \mathbf{x}) + M(\mathbf{x}, \mathbf{x}, \mathbf{x})\right) dt + a(\mathbf{x}, \sigma) d\mathbf{W}, \tag{2}$$

where M denotes a cubic nonlinear term, W is the Wiener process and σ the diffusion parameters.

In this study we will develop a systematic Bayesian framework for the efficient estimation of the model parameters using Markov Chain Monte Carlo (MCMC) methods from partial observations. We are dealing with partial observations because we now only have knowledge of the few resolved modes $\mathbf{x}$ and are ignorant about the many unresolved modes $\mathbf{y}$.

Stochastic climate modeling is a complex problem and empirically estimating the parameters poses several problems. First, the nonlinearity of climate models requires an approximation of the likelihood function. While it can be shown that this approximation converges to the true likelihood, this is not necessarily the case for real world applications. Here we develop a MCMC algorithm for the first time for stochastic climate models and demonstrate that this algorithm performs well. Second, the nonlinearity of the problem causes the space of parameters leading to stable and physical meaningful solutions to become small as the dimension of the problem increases. We show that a lot of the posterior mass is on parameter values which lead to solutions exploding to infinity in finite time. To solve this problem we derive global stability conditions. These conditions take the form of a negative definite matrix. Hence, we devise a novel sampling strategy based on sampling non-negative matrices. We show that this sampling strategy is computationally efficient and leads to stable solutions.

In section 2 we introduce stochastic climate models and derive conditions for global stability. Previous studies have shown that reduced climate models with quadratic nonlinearity experience unphysical finite time blow up and long time instabilities (19, 27, 28, 41). Here we use the normal form of stochastic climate models (26) and derive sufficient conditions for global stability for the normal form of stochastic climate models. These normal form stochastic climate models have cubic nonlinearities. The here derived stability condition is more general than the one in Majda et al. (26). In section 3 we develop a Bayesian framework for the systematic estimation of the model parameters using physical constraints. Here we develop an efficient way of sampling negative-definite matrices. Without this constraint the MCMC algorithm would produce about 40% unphysical solutions which is clearly very inefficient. Here we also demonstrate that for these kinds of SDEs imputing data improves the parameter estimates considerably. In section 4 we demonstrate the accuracy of our framework on conceptual climate models. We summarize our results in section 5.

2. Stochastic Climate Models and Global Stability

Here we study the following D dimensional normal form of stochastic climate models (which has the same structural form as Eq. (2), see (26)):

$$dx_i = \left(\alpha_i + \sum_{j=1}^{D} \beta_{i,j} x_j + \sum_{j+1}^{D} \sum_{k=1}^{j} \gamma_{i,j,k} x_j x_k + \sum_{j=1}^{D} \sum_{k=1}^{j} \sum_{l=1}^{k} \lambda_{i,j,k,l} x_j x_k x_l \right) dt \tag{3a}$$

$$+ \sum_{j=1}^{D} a_{i,j} dW_j + \sum_{j=1}^{D} \sum_{k=1}^{j} b_{i,j,k} x_j dW_k. \tag{3b}$$

which we write for convenience in a more compact form

$$d\mathbf{x} = \mu(\mathbf{x}, \mathbf{A})dt + \mathbf{a}(\mathbf{x}, \boldsymbol{\sigma})dW. \tag{4}$$

The parameters α, β, γ and λ are written as one matrix $\mathbf{A} \in \mathbb{R}^{D \times P}$. We allow for the inclusion of all possible linear, quadratic and cubic terms with forcing terms entering first, followed by linear, quadratic and then the cubic terms. We include them into a matrix $\mathbf{A}$ as $A_{i,1} = \alpha_i$, $A_{i,j+1} = \beta_{i,j}$, $A_{i,j(j-1)/2+k+D+1} = \gamma_{i,j,k}$ and $A_{i,f(j,k,l)} = \lambda_{i,j,k,l}$. The index function for the cubic term is given by

$$f(j,k,l) = 1 + D + \frac{D(D+1)}{2} + \frac{j(j-1)(j+1)}{6} + \frac{k(k-1)}{2} + l. \tag{5}$$

Global stability implies that the cubic terms act as a nonlinear damping. For climate models energy conservation of the nonlinear operator implies global stability. However, in general, global stability does not necessarily imply energy conservation. The energy equation based only on the cubic terms can be written as

$$\frac{1}{2}\frac{dE}{dt} = \sum_{i=1}^{D} x_i \frac{dx_i}{dt} = \sum_{i=1}^{D} \sum_{j=1}^{D} \sum_{k=1}^{j} \sum_{l=1}^{k} A_{i,f(j,k,l)} x_i x_j x_k x_l. \tag{6}$$

Here we only consider the cubic term because this term will ultimately determine global stability. Majda et al. (26) have shown that the normal form of stochastic climate models allows for linearly unstable modes. These linearly unstable modes are associated with important weather systems and waves and are an intrinsic and important part of climate models. Once these linearly unstable modes have reached a certain amplitude the nonlinear cubic terms will govern their evolution and, thus, ensure global stability.

We consider now the vector **v** with $\frac{D(D+1)}{2}$ components of the form $v_{(i-1)i/2+j} = x_i x_j$ with $1 \leq j \leq i \leq D$. Now we find a negative definite matrix **M** such that (26)

$$\mathbf{v}^T \mathbf{M} \mathbf{v} = \frac{1}{2}\frac{dE}{dt} \leq 0. \tag{7}$$

A sufficient solution is as follows: Let matrix $M \in \mathbb{R}^{(D+1)D/2 \times (D+1)D/2}$ be

$$M_{(i-1)i/2+j,(k-1)k/2+l} = \begin{cases} A_{i,f(i,j,l)}, & \text{if k>j and l} \leq \text{j} \\ 0, & \text{if k> j and l> j} \\ A_{i,f(i,j,l)} + A_{i,f(i,j,k)}, & \text{if k} \leq \text{j and l<k} \\ A_{i,f(i,j,l)}, & \text{if k} \leq \text{j and l=k,} \end{cases} \tag{8}$$

where $1 \leq j \leq i \leq D$ and $1 \leq l \leq k \leq D$. While this solution is not necessarily unique the imposition of this constraint is still necessary in order to reduce the amount of parameter values leading to unstable models and, thus, to reduce the computational expense of the parameter inference. As we will show below, not imposing this constraint will lead to unstable and unphysical solutions in about 40% of parameter values in our MCMC scheme.

In summary, the stochastic climate model in Eq. (4) is globally stable if the tensor **M** is negative definite and M determines the components of the cubic operator λ. This is an important result for the constrained parameter estimation of stochastic climate models (26, 32).

3. Physics Constraint Parameter Sampling

We use a Markov Chain Monte Carlo algorithm for the parameter inference which was proposed by (7) and (18). We use this approach because of its flexibility. For instance, the exact algorithms by (5) and (6) cannot be easily applied to multidimensional diffusions. Furthermore, the exact algorithms require that the drift function must be the gradient of a potential; for instance, stochastic climate models cannot be written in such a form.

Our MCMC algorithm first updates the diffusion parameters (see algorithm 1), then updates the imputed data (algorithm 2) and finally updates the drift parameters (section 3.2). To efficiently propose imputed data we us the Modified Linear Bridge sampler (section 3.1). To physically constrain the drift parameters we develop a scheme to sample negative definite matrices (section 3.3 and algorithms 3 and 4).

The novel aspect of our MCMC algorithm is the physics constraint sampling which ensures the stability of the reduced stochastic model. Moreover, this algorithm overcomes the dependency between the diffusion parameters and the missing data by changing variables to the underlying Brownian motion $\boldsymbol{W} \in \mathbb{R}^d$ and conditioning on this when performing the parameter update. This ensures consistency between the parameters and the path (7, 18). In order to improve the accuracy of the Euler approximation we introduce latent data points between all pairs of observations. While this is not trivial for nonlinear models this can be accomplished by introducing a suitable diffusion bridge (10, 18). For this purpose we define a process $\boldsymbol{Z}$, which conditions on the endpoint $\boldsymbol{x}_T$, by

$$d\boldsymbol{X}_t = \boldsymbol{a}(\boldsymbol{X}_t, \boldsymbol{\sigma})d\boldsymbol{Z}_t + \frac{\boldsymbol{x}_T - \boldsymbol{X}_t}{T-t}dt, \;\; \boldsymbol{X}_0 = \boldsymbol{x}_0\,, \tag{9}$$

where T is the next observation time. In discrete time the transformation is

$$\boldsymbol{X}_{i+1} = \boldsymbol{X}_i + \boldsymbol{a}(\boldsymbol{X}_i, \boldsymbol{\sigma})(\boldsymbol{Z}_{i+1} - \boldsymbol{Z}_i) + \frac{\boldsymbol{x}_T - \boldsymbol{X}_i}{m-i} \tag{10}$$

where $m-1$ is the number of imputed points between two observations. Defining the process $\boldsymbol{Z}$ ensures that the dominating measure is parameter free and, hence, improves the performance of the MH sampler. See Dargatz (9) for more details.

We sample $\boldsymbol{\sigma}$ according to Algorithm 1. We use zero-based numbering and $N-1$ observation intervals indexed $0 \dots N-2$. We assume that the inter-observation times Δ are all equal and that there are $m-1$ imputed points per interval, giving a time interval of $\delta = \Delta/m$. We use the notation $\boldsymbol{X}_i = \boldsymbol{X}_{t_i}$ and $\boldsymbol{\mu}_i = \boldsymbol{\mu}(\boldsymbol{X}_{t_i}, \boldsymbol{A})$. We assume that we have perfect observations for ease of notation but our method can be extended to the case of measurement error (e.g. (18)). The extension to variable inter-observation times is straight forward. For simplicity, we write the algorithm for perfect observation of the system so that $\boldsymbol{X}_{im}, i = 0, \dots, N-1$ are fixed. In Algorithm 1, ϕ denotes a Gaussian distribution and q the Gaussian proposal density (which is defined below in Eq. 19) and Σ a covariance matrix (given below in Eq. 16).

Algorithm 1 Sample parameters entering the diffusion matrix.

```
Draw $\boldsymbol{\sigma}^* \sim q(\boldsymbol{\sigma}^*|\boldsymbol{\sigma})$
Initialize $\alpha = \log(q(\boldsymbol{\sigma}|\boldsymbol{\sigma}^*)) - \log(q(\boldsymbol{\sigma}^*|\boldsymbol{\sigma})) + \log(p(\boldsymbol{\sigma}^*)) - \log(p(\boldsymbol{\sigma}))$
for $i = 0$ to $N-2$ do
  for $j = 0$ to $m-2$ do
    $\boldsymbol{Z}_{im+j+1} = \boldsymbol{Z}_{im+j} + \boldsymbol{a}^{-1}(\boldsymbol{X}_{im+j}, \boldsymbol{\sigma})\left(\boldsymbol{X}_{im+j+1} - \boldsymbol{X}_{im+j} - \frac{\boldsymbol{X}_{im+m} - \boldsymbol{X}_{im+j}}{m-j}\right)$
    $\boldsymbol{X}^*_{im+j+1} = \boldsymbol{X}^*_{im+j} + \frac{\boldsymbol{X}_{im+m} - \boldsymbol{X}^*_{im+j}}{m-j} + \boldsymbol{a}(\boldsymbol{X}^*_{im+j}, \boldsymbol{\sigma}^*)(\boldsymbol{Z}_{im+j+1} - \boldsymbol{Z}_{im+j})$
    $\alpha = \alpha + \log(\phi(\boldsymbol{X}^*_{im+j+1}; \boldsymbol{X}^*_{im+j} + \boldsymbol{\mu}^*_{im+j}\delta, \delta\boldsymbol{\Sigma}^*_{im+j}) + \log|a(\boldsymbol{X}^*_{im+j}, \boldsymbol{\sigma}^*)|$
        $- \log(\phi(\boldsymbol{X}_{im+j+1}; \boldsymbol{X}_{im+j} + \boldsymbol{\mu}_{im+j}\delta, \delta\boldsymbol{\Sigma}_{im+j}) - \log|a(\boldsymbol{X}_{im+j}, \boldsymbol{\sigma})|$
  end for
end for
Set $\{\boldsymbol{\sigma}, \boldsymbol{X}\} = \{\boldsymbol{\sigma}^*, \boldsymbol{X}^*\}$ with probability $\min(1, \exp(\alpha))$ else retain $\{\boldsymbol{\sigma}, \boldsymbol{X}\}$
```

To update missing data between observations we use an independence sampler as in (36) using the proposal process

$$d\boldsymbol{X}^* = \boldsymbol{\xi}(\boldsymbol{X}^*, \boldsymbol{X}_T)dt + \boldsymbol{a}(\boldsymbol{X}^*, \boldsymbol{\sigma})d\boldsymbol{W}^*\;, \tag{11}$$

where $\boldsymbol{X}_T$ is the next observation, $\boldsymbol{X}^*$ the proposed data, and where $\boldsymbol{a}(\boldsymbol{X}^*, \boldsymbol{\sigma})$ is the same diffusion function as that in Eq. (4). $\boldsymbol{\xi}$ denotes the modified linear bridge (see below in section 3.1). The proposal process Eq. (11) will have a measure that is absolutely continuous with respect to the target process in Eq. (4) because of their common diffusion function. To update all of the missing data, we propose a block at a time from Eq. (11) and then accept the proposed block according to the MH ratio. If the inter-observation interval is large then the acceptance rate may become very low and so one may sub-sample smaller blocks.

For some interval i we set $\boldsymbol{X}^*_0 = \boldsymbol{X}_{im}$ and $\boldsymbol{X}^*_m = \boldsymbol{X}_{(i+1)m}$ then we propose $\boldsymbol{X}^*_1 : \boldsymbol{X}^*_{m-1}$ and accept or reject the block using the MH acceptance probability

$$\alpha = \frac{p_\delta(\boldsymbol{X}^*_m|\boldsymbol{X}^*_{m-1}, \boldsymbol{A}) \prod_{j=0}^{m-2} p_\delta(\boldsymbol{X}^*_{j+1}|\boldsymbol{X}^*_j, \boldsymbol{A}) q_\delta(\boldsymbol{X}_{im+j+1}|\boldsymbol{X}_{im+j}, \boldsymbol{\xi}, \boldsymbol{\sigma})}{p_\delta(\boldsymbol{X}_{(i+1)m}|\boldsymbol{X}_{im+m-1}, \boldsymbol{A}) \prod_{j=0}^{m-2} p_\delta(\boldsymbol{X}_{im+j+1}|\boldsymbol{X}_{im+j}, \boldsymbol{A}) q_\delta(\boldsymbol{X}^*_{j+1}|\boldsymbol{X}^*_j, \boldsymbol{\xi}, \boldsymbol{\sigma})}, \tag{12}$$

where p_δ is the transition density of the target

$$d\boldsymbol{X}_t = \boldsymbol{\mu}(\boldsymbol{X}_t, \boldsymbol{A})dt + \boldsymbol{a}(\boldsymbol{X}_t, \boldsymbol{\sigma})d\boldsymbol{W}_t\,,\ \ \boldsymbol{X}_0 = \boldsymbol{x}_0\,,\ \ t \in [0, T] \tag{13}$$

under the Euler approximation over the time interval δ and where q_δ denotes the transition density of the proposal. We choose proposal processes so that given $\boldsymbol{X}^*_j$, $\boldsymbol{X}^*_{j+1}$ is approximately Gaussian distributed. However, Eq. (11) is not a true Gaussian process because of the state dependent noise term. Details for updating the missing data are given in Algorithm 2.

Algorithm 2 Sample missing data between observations.

```
for $i = 0$ to $N - 2$ do
    Set $\boldsymbol{X}^*_0 = \boldsymbol{X}_{im}$
    Set $\alpha = 0$
    for $j = 0$ to $m - 2$ do
        $\boldsymbol{X}^*_{j+1} \sim q_\delta(\boldsymbol{X}^*_{j+1}|\boldsymbol{\xi}(\boldsymbol{X}^*_j, \boldsymbol{X}_{im+m}), \boldsymbol{\sigma})$
            $\alpha = \alpha + \log(\phi(\boldsymbol{X}^*_{j+1}; \boldsymbol{X}^*_j + \delta\boldsymbol{\mu}^*_j, \delta\boldsymbol{\Sigma}^*_j) + \log(q_\delta(\boldsymbol{X}_{im+j+1}|\boldsymbol{\xi}(\boldsymbol{X}_{im+j}, \boldsymbol{X}_{im+m}), \boldsymbol{\sigma}))$
            $- \log(\phi(\boldsymbol{X}_{im+j+1}; \boldsymbol{X}_{im+j} + \delta\boldsymbol{\mu}_{im+j}, \delta\boldsymbol{\Sigma}_{im+j})) - \log(q_\delta(\boldsymbol{X}^*_{j+1}|\boldsymbol{X}^*_j, \boldsymbol{\xi}(\boldsymbol{X}^*_j, \boldsymbol{X}_{im+m}), \boldsymbol{\sigma}))$
    end for
    $\alpha = \alpha + \log(\phi(\boldsymbol{X}_{im+m}; \boldsymbol{X}^*_{m-1} + \delta\boldsymbol{\mu}^*_{m-1}, \delta\boldsymbol{\Sigma}^*_{m-1}))$
    $- \log(\phi(\boldsymbol{X}_{im+m}; \boldsymbol{X}_{im+m-1} + \delta\boldsymbol{\mu}_{im+m-1}, \delta\boldsymbol{\Sigma}_{im+m-1}))$
    if $\exp(\alpha) > \mathcal{U}(0, 1)$ then
        for $j = 0$ to $m - 2$ do
            $\boldsymbol{X}_{im+j+1} = \boldsymbol{X}^*_{j+1}$
        end for
    end if
end for
```

Algorithms 1 and 2 are combined with standard MH updates for the parameters $\boldsymbol{A}$ entering into the drift function. First we update the diffusion parameters using Algorithm 1, then we update all imputed data. After that the drift parameters will be updated (see section 3.2). In both algorithms X^* denotes the proposal which will be generated dependending on the availability of observations. One could use Random-Walk proposals but in our case of polynomial models it is more efficient to implement another Gibbs sampling step. Repeatedly alternating between these three steps will produce MCMC samples that can be used to estimate the parameters. In practice we increase the amount of missing data m until we see convergence in the marginal distributions of the parameters.

3.1. Sampling of Diffusion Paths

Because we want to impute missing data we need efficient methods for simulating diffusion paths from Eq. (3) that are conditioned upon given start $\boldsymbol{X}_0 = \boldsymbol{x}_0$ and end $\boldsymbol{X}_m = \boldsymbol{x}_m$ points. We consider the total time interval $\tau_m - \tau_0 = \Delta$ divided into m equidistant sub-intervals so that $\tau_{k+1} - \tau_k = \Delta/m = \delta$.

Having N observations, for each of $i = 0, 1, 2, \ldots N-1$, X_{im} is an observation. Between every pair of observations the diffusion bridge will need to be simulated. We use an independence sampler with proposal density of the form $q(\boldsymbol{X}^*|\boldsymbol{X}) = q(\boldsymbol{X}^*)$. Here, we consider proposal processes of the form of Eq. (11).

We use a Modified Linear Bridge proposal for sampling of parameters of the drift of equations of the form of Eq. 4. For this purpose we apply Ito's formula to the drift function of Eq. 4. This gives the approximating process

$$d\boldsymbol{Z}_t = (\boldsymbol{Q}(\boldsymbol{X})\boldsymbol{Z}_t + \boldsymbol{r}(\boldsymbol{X}, t))dt + \boldsymbol{\Sigma}(\boldsymbol{X})d\boldsymbol{B}_t. \tag{14}$$

with

$$\begin{aligned}
Q_{ij} &= \frac{\partial \mu_i(\boldsymbol{X}_s)}{\partial x_j} \\
r_i(t) &= \mu_i(\boldsymbol{X}_s) - \sum_j \frac{\partial \mu_i}{\partial x_j} \boldsymbol{X}_j(s) + \frac{1}{2} \sum_{j,k,l} a_{jl}(\boldsymbol{X}_t) a_{kl}(\boldsymbol{X}_t) \frac{\partial^2 \mu_i}{\partial x_j \partial x_k}(\boldsymbol{X}_t)(t-s) \\
\boldsymbol{\Sigma} &= \boldsymbol{a}(\boldsymbol{X}_s)
\end{aligned}$$

This is a local linearization of the nonlinear diffusion over a small time window (30, 39). First we construct bridge distributions for general multivariate linear diffusions (4). If at time s we have $\boldsymbol{X}_s = \boldsymbol{d}$ and at time T, $\boldsymbol{X}_T = \boldsymbol{e}$ then the distribution of $\boldsymbol{X}_t$ for $0 \leq s < t \leq T$ can be shown to be Gaussian with mean

$$\boldsymbol{\nu}_{\boldsymbol{d},\boldsymbol{e}}(s,t) = \boldsymbol{\Gamma}(t,T)\boldsymbol{\Gamma}(s,T)^{-1}\boldsymbol{m}^{+}_{\boldsymbol{d}}(s,t) + \boldsymbol{\Gamma}(s,T)^{T}(\boldsymbol{\Gamma}(s,T)^{T})^{-1}\boldsymbol{m}^{-}_{\boldsymbol{e}}(t,T), \tag{15}$$

where

$$\boldsymbol{\Gamma}(s,t) = \int_s^t e^{(s-u)\boldsymbol{Q}} \boldsymbol{\Sigma}\boldsymbol{\Sigma}^T e^{(t-u)\boldsymbol{Q}^T} du,$$

$$\boldsymbol{m}^{+}_{\boldsymbol{x}}(s,t) = \boldsymbol{x} + \int_s^t e^{(s-u)\boldsymbol{Q}} \boldsymbol{r}(u) du \quad \text{and} \quad \boldsymbol{m}^{-}_{\boldsymbol{x}}(s,t) = \boldsymbol{x} - \int_s^t e^{(t-u)\boldsymbol{Q}} \boldsymbol{r}(u) du.$$

The covariance matrix is given by

$$\boldsymbol{\Sigma}(s,t) = \boldsymbol{\Gamma}(t,T)\boldsymbol{\Gamma}(s,T)^{-1}\boldsymbol{\Gamma}(s,t). \tag{16}$$

In general this matrix can be computed as follows: if we diagonalize $\boldsymbol{Q}$ so that $\boldsymbol{Q} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{-1}$ then compute the matrix $\boldsymbol{A}$ with components

$$V_{ij} = \frac{(\boldsymbol{U}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^T\boldsymbol{U}^{-T})_{ij}}{\Lambda_{ii} + \Lambda_{jj}} \left(e^{(t-s)\Lambda_{jj}} - e^{(s-t)\Lambda_{ii}}\right), \tag{17}$$

then

$$\boldsymbol{\Gamma}(s,t) = \boldsymbol{U}\boldsymbol{V}\boldsymbol{U}^T. \tag{18}$$

The proposal distribution is given by

$$q(\boldsymbol{X}_{k+1}|\boldsymbol{X}_k, \boldsymbol{X}_m, \boldsymbol{V}) = \phi\left(\boldsymbol{X}_{k+1}; \boldsymbol{\nu}_{\boldsymbol{x}_k,\boldsymbol{x}_m}(k\delta,(k+1)\delta), \boldsymbol{\Sigma}_{\boldsymbol{x}_k}(k\delta,(k+1)\delta)\right), \tag{19}$$

where $\boldsymbol{\nu}_{\boldsymbol{x}_j,\boldsymbol{x}_m}(j\delta,(j+1)\delta)$ and $\boldsymbol{\Sigma}_{\boldsymbol{x}_0}(j\delta,(j+1)\delta)$ are given in Eqs. (15) and (16) respectively.

In contrast to a linear bridge sampler here we update at each imputed point. This means recomputing the matrices $\boldsymbol{\Gamma}(s,t)$ at each point, although $\boldsymbol{Q}$ and $\boldsymbol{U}$ are only calculated once.

3.2. Inference for Drift Parameters

Now we give details of the computational implementation of the sampling of parameters in the drift function. Since the drift parameters enter linearly we can construct a Gibbs sampler where their conditional posterior is Gaussian. This greatly improves the mixing of the Markov Chain.

3.2.1. Gibbs Sampler

Consider N observations with time interval δ. We set $\boldsymbol{Y}_t = \boldsymbol{X}_{t+1} - \boldsymbol{X}_t$ and let $\boldsymbol{U} \in \mathbb{R}^{N-1\times P}$ be the design matrix of the data, scaled by δ. The columns of $\boldsymbol{U}$ are indexed in the same way as the columns of the parameter matrix $\boldsymbol{A}$. For example, a two dimensional system would have $P = 10$ and the following design matrix

$$U = \delta \left(\begin{matrix} 1 & X_{1,1} & X_{1,2} & X_{1,1}^2 & X_{1,1}X_{1,2} & X_{1,2}^2 & X_{1,1}^3 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{N-1,1} & X_{N-1,2} & X_{N-1,1}^2 & X_{N-1,1}X_{N-1,2} & X_{N-1,2}^2 & X_{N-1,1}^3 \end{matrix} \right.$$

$$\left. \begin{matrix} X_{1,1}^2 X_{1,2} & X_{1,1}X_{1,2}^2 & X_{1,2}^3 \\ \vdots & \vdots & \vdots \\ X_{N-1,1}^2 X_{N-1,2} & X_{N-1,1}X_{N-1,2}^2 & X_{N-1,2}^3 \end{matrix} \right)$$

The log likelihood can be written

$$L(\boldsymbol{A};\boldsymbol{X}) = -\frac{1}{2}\sum_{t=1}^{N-1} |\boldsymbol{\Sigma}_{Drift\ t}| - \frac{1}{2}\sum_{t=1}^{N-1}\sum_{i,j=1}^{D}\left(\boldsymbol{Y}_{ti} - \sum_{k=1}^{P} U_{tk}A_{ik}\right)\Sigma_{Drift\ tij}^{-1}\left(Y_{tj} - \sum_{k=1}^{P} U_{tk}A_{jk}\right), \tag{20}$$

where the instantaneous covariance matrix $\boldsymbol{\Sigma}_{Drift\ t}$ is computed from $\Sigma_{Drift\ t,j,k}^{1/2} = (d_{j,k} + \sum_{l=1}^{D} e_{l,j,k}X_{t,j})\Delta^{1/2}$.

We have DP parameters to infer in the matrix $\boldsymbol{A}$. We use a zero mean Gaussian prior with covariance matrix $\boldsymbol{\Gamma}_{\boldsymbol{Drift}} \in \mathbb{R}^{DP\times DP}$. Let $\Lambda \in \mathbb{R}^{DP\times DP}$, be a matrix with components

$$\Lambda_{(i-1)P+j,(k-1)P+l} = \sum_{t=1}^{N-1} U_{tj}\Sigma_{Drift\ tik}^{-1}U_{tl} + \Gamma_{Drift\ (i-1)P+j,(k-1)P+l}^{-1} \tag{21}$$

where $i,k = 1\ldots D$ and $j,l = 1\ldots P$. Let $e \in \mathbb{R}^{DP}$ with components

$$e_{(i-1)P+j} = \sum_{t,k} U_{t,j}\Sigma_{Drift\ tik}^{-1}Y_{tk}. \tag{22}$$

The posterior mean $\mu_{(i-1)P+j}$ of $A_{i,j}$ is given by the solution of $\boldsymbol{\Lambda\mu} = \boldsymbol{b}$ and the posterior covariance is $\mathrm{Cov}(A_{i,j}, A_{k,l}) = \Lambda_{(i-1)P+j,(k-1)P+l}^{-1}$.

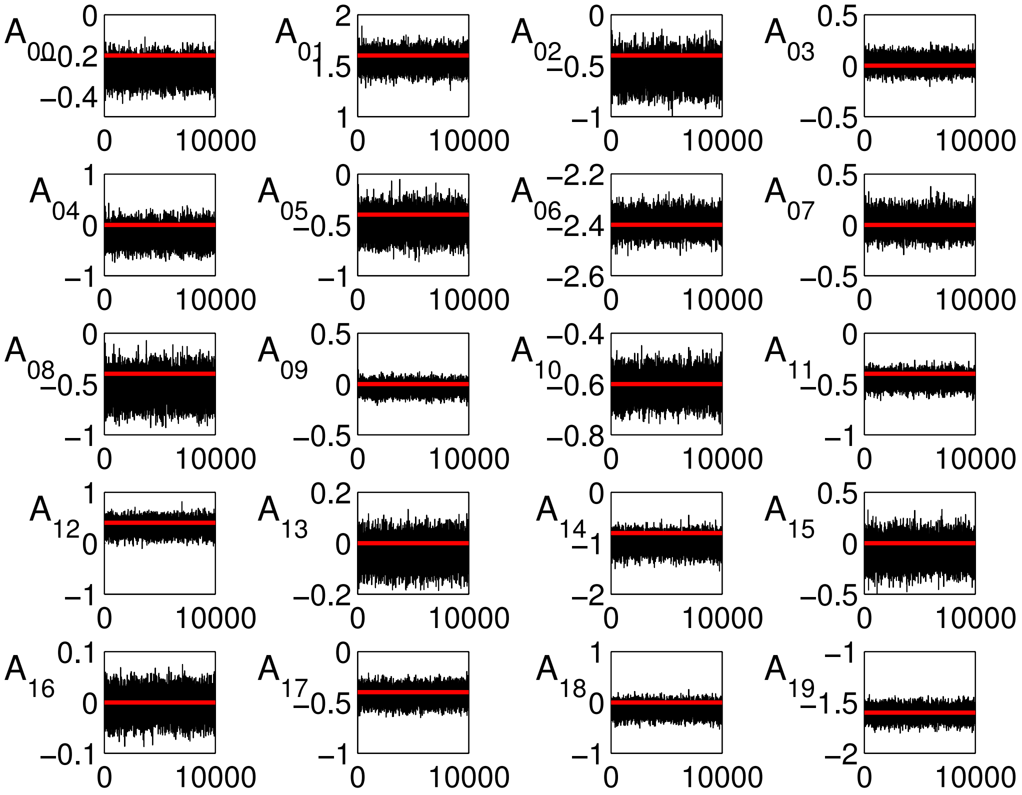


Figure 1: Output of a Gibbs sampler for 20 drift parameters of a two dimensional model from Eq. (3). The observation interval is $\delta = 10^{-3}$ and $T = 10,000$. The true values are shown in red.

We applied the above Gibbs sampler to a large data set from a two dimensional model of the form of Eq. (3) with random values for the diffusion parameters. We chose a fine observation interval of $\delta = 10^{-3}$ and long observation period $T = 10,000$. Fig. 1 displays the trace plots for all 20 parameters (note that the indices are from 0 rather than 1 as in the text). Using this large data set the algorithm is able to reproduce the true values shown in red.

We performed a further simulation study to test the dependence of the posterior estimates upon the data set used. We inferred all of the drift parameters for a simple two dimensional model of the form of Eq. (3) using data sets of length $T = \{10, 100, 1000\}$ and with observation interval $\Delta = \{0.1, 0.01, 0.001\}$. Note that the diffusion function is arbitrary in this model. The results are shown in Table 1. For each parameter we estimated the posterior mean and the posterior 10th-90th percentiles. A cell is colored blue where the true value falls within this range. Note that if we were using the true likelihood, rather than an approximation, then we would expect there to be around 80% blue boxes. The error of the estimates for each data set can be quantified using the quadratic **Posterior Expected Loss** (PEL) function

$$f(\hat{\pi}, \boldsymbol{X}_{\text{obs}}) = \int_{\Theta} (\boldsymbol{\theta}^* - \boldsymbol{\theta})^2 \hat{\pi}(\boldsymbol{\theta}|\boldsymbol{X}_{\text{obs}}) d\boldsymbol{\theta}\,, \tag{23}$$

where $\boldsymbol{\theta}$ represents all of the parameters, $\hat{\pi}$ is the estimated posterior distribution and $\boldsymbol{\theta}^*$ is the true value of the parameter.

We performed a test with both the Gibbs sampler and data imputation. In Tab. 2 the data is observed at interval $\Delta = 0.1$. The smaller intervals $\Delta = \{0.01, 0.001\}$ are obtained by imputing data with $m = \{10, 100\}$ respectively. The table shows that imputing data approximately doubles the Posterior Expected Loss. As expected the confidence intervals are broader but with more imputed data the algorithm can recover the true values. This shows that our data imputing strategy successfully improves the parameter estimates.

Our aim is to infer models that can be used for prediction. This can be problematic when dealing with non-linear models as some (generally unknown) regions of the parameter space will give solutions that explode to infinity with probability 1. This is a particular problem when, as exemplified by Table 1, large amounts of data are needed to regain the true values.

To demonstrate this problem we performed an inference on a two dimensional cubic model using $N = 1,000$ observations at $\Delta = 0.1$. For each inferred parameter value we then simulated the solution for $T = 100$. After this time we recorded whether the solution retained finite values or had exploded. The marginal posterior distributions of the cubic parameters are plotted in Fig. 2. Each plot shows two histograms: one in blue records the distribution of stable parameter values and in red are those that exploded. Notice that, when looking at the marginal distributions, the stable and unstable regions largely overlap; it is difficult to separate the two regions. In this case 40% of values were unstable. Tests (not shown) indicate that this is an even bigger problem in higher dimensions. Therefore, it is essential to use constraints on the parameter space to enable only physically meaningful solutions. The necessary conditions have been derived in section 2.

Thus, as shown above, when updating the drift parameters we ensure that $\boldsymbol{M}$ is negative definite. In practice it is sufficient to check only whether the symmetric part $(\boldsymbol{M} + \boldsymbol{M}^T)/2$ is negative-definite. In the next section we will develop a systematic way of sampling negative definite matrices.

3.3. Sampling Negative Definite Matrices

To sample negative definite matrices we use the Component Wise algorithm (32). Here we sample the density of a $n \times n$ matrix $\boldsymbol{M}$, with normally distributed components, subject to the constraint that it is negative definite. This algorithm updates $\boldsymbol{M}$ component wise and is based on the following property: a $n \times n$ matrix is negative definite if and only if all $k \leq n$ leading principal minors obey $|\boldsymbol{M}^{(k)}|(-1)^k > 0$. The kth principal minor is the determinant of the upper left $k \times k$ sub-matrix. Consider the parameters along the main diagonal. As they only enter $\boldsymbol{M}$ once, each will have an associated upper bound. The Algorithm 3 works by calculating the upper bound associated with the constraints from each principal minor. It does this to find the least upper bound and thereby the truncation point of the normal distribution.

Here, $\mathcal{N}_-(\mu, u, \sigma^2)$ is the right truncated normal distribution with mean μ, standard deviation σ and upper bound u. The off-diagonal parameters enter twice so there will be a quadratic function determining their limits for each leading principal minor. For parameters in element $M_{ij}^{(k)}$ there will be an associated quadratic relation $a_{ij}^{(k)} M_{ij}^2 + b_{ij}^{(k)} M_{ij} + c_{ij}^{(k)} = 0$ where the coefficients

	$T = 10$			$T = 100$			$T = 1000$		
	0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001
$A_{00} = 0$	2.18 (−2.5,6.89)	−0.44 (−4.79,3.74)	−1.95 (−6.07,2.2)	−0.13 (−1.07,0.77)	0.65 (−0.29,1.6)	0.64 (−0.28,1.58)	0.02 (−0.13,0.17)	0.01 (−0.13,0.17)	0.02 (−0.13,0.18)
$A_{01} = 5$	−2.04 (−7.23,3.04)	4.86 (0.84,9.04)	6.21 (2.31,10.12)	2.54 (1.78,3.28)	4.73 (3.93,5.51)	5.22 (4.43,6.03)	2.84 (2.7,2.97)	4.59 (4.45,4.73)	4.86 (4.73,5)
$A_{02} = 0$	2.63 (−0.16,5.39)	2.15 (−0.17,4.49)	1.63 (−0.66,4)	0.45 (−0.52,1.41)	0.58 (−0.42,1.6)	0.13 (−0.89,1.13)	−0.08 (−0.22,0.05)	−0.14 (−0.27,0)	−0.19 (−0.32,−0.04)
$A_{03} = 0$	1.54 (−2.75,5.84)	−0.17 (−3.12,2.76)	0.26 (−2.62,3.2)	0.27 (−0.38,0.93)	0.01 (−0.72,0.73)	−0.17 (−0.88,0.54)	−0.01 (−0.08,0.06)	0 (−0.07,0.06)	0 (−0.07,0.06)
$A_{04} = 0$	−3.51 (−8.17,1.13)	0.61 (−3,4.3)	1.3 (−2.32,5.07)	−0.96 (−2.09,0.2)	−0.58 (−1.8,0.68)	−0.07 (−1.3,1.16)	−0.01 (−0.04,0.01)	−0.02 (−0.04,0.01)	−0.02 (−0.05,0.01)
$A_{05} = 0$	0.32 (−2.19,2.81)	0 (−2.31,2.31)	0.19 (−2.11,2.49)	0.66 (−0.11,1.43)	0.02 (−0.79,0.84)	−0.27 (−1.05,0.55)	0 (−0.06,0.07)	0.01 (−0.06,0.07)	0 (−0.07,0.07)
$A_{06} = -3$	−0.88 (−2.96,1.24)	−2.58 (−4.3,−0.87)	−3.17 (−4.83,−1.49)	−1.8 (−2,−1.6)	−2.92 (−3.12,−2.71)	−3.05 (−3.25,−2.85)	−1.71 (−1.77,−1.65)	−2.75 (−2.81,−2.69)	−2.91 (−2.97,−2.85)
$A_{07} = 0$	0.58 (−1.87,3.06)	−1.32 (−3.32,0.7)	−1.48 (−3.54,0.52)	0.05 (−0.49,0.58)	−0.07 (−0.66,0.51)	−0.22 (−0.79,0.37)	−0.02 (−0.07,0.04)	−0.01 (−0.06,0.05)	0 (−0.05,0.06)
$A_{08} = 0$	−0.83 (−3,1.31)	−0.28 (−2.31,1.76)	−0.18 (−2.21,1.86)	−0.48 (−1.03,0.07)	−0.37 (−0.96,0.23)	−0.16 (−0.75,0.42)	−0.02 (−0.08,0.03)	−0.01 (−0.06,0.04)	−0.01 (−0.07,0.04)
$A_{09} = 0$	0.71 (−0.11,1.53)	−0.4 (−1.16,0.38)	−0.49 (−1.26,0.27)	0.33 (−0.02,0.67)	0.04 (−0.3,0.39)	0.09 (−0.25,0.44)	0.08 (0.02,0.13)	0.1 (0.04,0.16)	0.12 (0.05,0.17)
$A_{10} = 0$	−0.38 (−5.19,4.24)	−2.21 (−6.44,2.03)	−1.17 (−5.21,2.9)	−0.31 (−1.22,0.61)	−0.18 (−1.1,0.75)	0.06 (−0.89,0.99)	0.02 (−0.13,0.17)	0.05 (−0.1,0.2)	0.05 (−0.1,0.2)
$A_{11} = 0$	−2.93 (−8.11,2.15)	2.42 (−1.64,6.48)	1.67 (−2.29,5.64)	0.88 (0.1,1.64)	0.39 (−0.4,1.19)	0.33 (−0.48,1.12)	−0.06 (−0.2,0.08)	0.11 (−0.03,0.25)	0.11 (−0.04,0.24)
$A_{12} = 5$	5.96 (3.22,8.68)	4.2 (1.89,6.63)	4.72 (2.33,7.09)	1.68 (0.72,2.65)	4.56 (3.52,5.57)	5.09 (4.08,6.12)	2.88 (2.74,3.02)	4.7 (4.56,4.84)	4.98 (4.84,5.11)
$A_{13} = 0$	2.68 (−1.67,6.93)	−1.41 (−4.36,1.57)	−1.88 (−4.74,1.02)	−1.05 (−1.72,−0.38)	−0.61 (−1.33,0.1)	−0.74 (−1.46,−0.02)	−0.01 (−0.08,0.05)	−0.03 (−0.1,0.03)	−0.03 (−0.09,0.04)
$A_{14} = 0$	−1.13 (−5.84,3.6)	4.67 (1.01,8.45)	4.23 (0.47,7.92)	1.39 (0.24,2.55)	0.78 (−0.44,2.03)	0.91 (−0.33,2.14)	0.02 (−0.01,0.05)	0.03 (0,0.05)	0.03 (0,0.05)
$A_{15} = 0$	0.42 (−2.07,2.94)	−0.6 (−2.88,1.77)	−0.84 (−3.13,1.45)	−0.22 (−1.01,0.56)	−0.11 (−0.9,0.7)	−0.24 (−1.06,0.55)	0 (−0.07,0.06)	−0.01 (−0.08,0.05)	−0.02 (−0.09,0.04)
$A_{16} = 0$	−0.82 (−2.94,1.29)	0.12 (−1.61,1.83)	0.39 (−1.26,2)	−0.08 (−0.28,0.12)	−0.04 (−0.24,0.17)	−0.02 (−0.22,0.19)	−0.03 (−0.09,0.03)	−0.08 (−0.14,−0.02)	−0.09 (−0.15,−0.03)
$A_{17} = 0$	−0.46 (−2.95,2.04)	−3.21 (−5.28,−1.22)	−2.97 (−4.99,−0.93)	−0.71 (−1.25,−0.17)	−0.66 (−1.24,−0.08)	−0.78 (−1.35,−0.19)	0.01 (−0.05,0.06)	−0.01 (−0.06,0.04)	0 (−0.05,0.05)
$A_{18} = 0$	0.66 (−1.5,2.86)	1.33 (−0.71,3.35)	1.63 (−0.41,3.6)	0.65 (0.11,1.22)	0.45 (−0.13,1.04)	0.58 (−0.01,1.18)	0.06 (0.01,0.12)	0.02 (−0.03,0.08)	0.03 (−0.02,0.09)
$A_{19} = -3$	−2.61 (−3.44,−1.78)	−3.23 (−4.02,−2.45)	−3.47 (−4.24,−2.69)	−1.46 (−1.8,−1.12)	−2.73 (−3.07,−2.38)	−3.04 (−3.39,−2.7)	−1.75 (−1.81,−1.69)	−2.81 (−2.87,−2.75)	−2.98 (−3.04,−2.92)
	8.48	5.25	4.96	1.19	0.37	0.36	0.43	0.02	0.01

Table 1: Drift parameter estimates for a two dimensional cubic model with arbitrary diffusion function. On the left is the true value of the parameter. The length of the data set used for the inference is labeled as T and the observation interval is $\Delta = \{0.1, 0.01, 0.001\}$. In each cell the parameter is estimated from the posterior mean and in brackets is shown the 10th-90th percentiles of the posterior. The blue coloring is where the true value falls in this range. The bottom of the table shows the Posterior Expected Loss in each case.

	$T = 10$			$T = 100$			$T = 1000$		
	0.1	0.01	0.001	0.1	0.01	0.001	0.1	0.01	0.001
$A_{00} = 0$	2.18 (−2.5,6.89)	−2.46 (−8.46,3.52)	−2.59 (−8.35,3.24)	−0.13 (−1.07,0.77)	−0.68 (−1.84,0.5)	−0.69 (−1.85,0.49)	0.02 (−0.13,0.17)	−0.08 (−0.27,0.09)	−0.08 (−0.26,0.09)
$A_{01} = 5$	−2.04 (−7.23,3.04)	4.99 (−0.7,10.61)	4.84 (−0.49,10.22)	2.54 (1.78,3.28)	4.89 (3.83,5.97)	4.76 (3.74,5.79)	2.84 (2.7,2.97)	5.04 (4.87,5.21)	4.88 (4.71,5.03)
$A_{02} = 0$	2.63 (−0.16,5.39)	0.61 (−2.94,4.05)	0.39 (−2.95,3.75)	0.45 (−0.52,1.41)	−0.06 (−1.42,1.3)	−0.2 (−1.49,1.13)	−0.08 (−0.22,0.05)	−0.26 (−0.44,−0.08)	−0.28 (−0.45,−0.09)
$A_{03} = 0$	1.54 (−2.75,5.84)	3.25 (−1.23,8.07)	3.26 (−1.07,7.77)	0.27 (−0.38,0.93)	0.67 (−0.34,1.67)	0.59 (−0.41,1.59)	−0.01 (−0.08,0.06)	0.02 (−0.06,0.1)	0.02 (−0.06,0.09)
$A_{04} = 0$	−3.51 (−8.17,1.13)	−2.53 (−7.76,2.61)	−2.53 (−7.61,2.41)	−0.96 (−2.09,0.2)	−1.6 (−3.35,0.18)	−1.41 (−3.14,0.27)	−0.01 (−0.04,0.01)	−0.03 (−0.05,0)	−0.02 (−0.05,0)
$A_{05} = 0$	0.32 (−2.19,2.81)	1.85 (−1.6,5.2)	1.85 (−1.43,5.11)	0.66 (−0.11,1.43)	1.18 (0.03,2.32)	1.12 (−0.02,2.25)	0 (−0.06,0.07)	0.05 (−0.04,0.13)	0.05 (−0.04,0.13)
$A_{06} = -3$	−0.88 (−2.96,1.24)	−4.06 (−6.63,−1.76)	−4.02 (−6.5,−1.79)	−1.8 (−2,−1.6)	−3.19 (−3.44,−2.94)	−3.06 (−3.3,−2.83)	−1.71 (−1.77,−1.65)	−3.01 (−3.08,−2.94)	−2.92 (−2.98,−2.85)
$A_{07} = 0$	0.58 (−1.87,3.06)	0.44 (−2.27,3.17)	0.53 (−2.05,3.23)	0.05 (−0.49,0.58)	0.43 (−0.37,1.27)	0.39 (−0.42,1.19)	−0.02 (−0.07,0.04)	0.03 (−0.04,0.09)	0.03 (−0.03,0.09)
$A_{08} = 0$	−0.83 (−3,1.31)	−2.22 (−4.98,0.62)	−2.12 (−4.84,0.65)	−0.48 (−1.03,0.07)	−1 (−1.86,−0.14)	−0.92 (−1.76,−0.09)	−0.02 (−0.08,0.03)	−0.02 (−0.09,0.05)	−0.02 (−0.09,0.05)
$A_{09} = 0$	0.71 (−0.11,1.53)	1.29 (0.06,2.61)	1.33 (0.1,2.6)	0.33 (−0.02,0.67)	0.7 (0.17,1.2)	0.71 (0.19,1.2)	0.08 (0.02,0.13)	0.14 (0.05,0.23)	0.15 (0.06,0.23)
$A_{10} = 0$	−0.38 (−5.19,4.24)	−4.63 (−10.46,0.84)	−4.46 (−10.07,0.86)	−0.31 (−1.22,0.61)	0.66 (−0.55,1.87)	0.6 (−0.51,1.75)	0.02 (−0.13,0.17)	−0.02 (−0.2,0.16)	−0.01 (−0.2,0.16)
$A_{11} = 0$	−2.93 (−8.11,2.15)	2.18 (−3.8,8.3)	2.24 (−3.53,8.27)	0.88 (0.1,1.64)	0.1 (−0.89,1.09)	0.15 (−0.82,1.14)	−0.06 (−0.2,0.08)	0.08 (−0.09,0.26)	0.11 (−0.06,0.29)
$A_{12} = 5$	5.96 (3.22,8.68)	8.28 (4.62,11.9)	7.82 (4.39,11.36)	1.68 (0.72,2.65)	5.8 (4.52,7.07)	5.57 (4.36,6.81)	2.88 (2.74,3.02)	5.05 (4.88,5.22)	4.89 (4.72,5.05)
$A_{13} = 0$	2.68 (−1.67,6.93)	4.03 (−0.84,9.1)	4.08 (−0.63,8.96)	−1.05 (−1.72,−0.38)	−0.83 (−1.76,0.11)	−0.81 (−1.72,0.1)	−0.01 (−0.08,0.05)	−0.01 (−0.09,0.08)	−0.01 (−0.09,0.08)
$A_{14} = 0$	−1.13 (−5.84,3.6)	0.2 (−5.79,6.51)	0.02 (−5.85,6.12)	1.39 (0.24,2.55)	0.12 (−1.44,1.66)	0.07 (−1.43,1.61)	0.02 (−0.01,0.05)	0.03 (0,0.06)	0.03 (0,0.06)
$A_{15} = 0$	0.42 (−2.07,2.94)	1.49 (−1.84,5)	1.59 (−1.64,4.85)	−0.22 (−1.01,0.56)	0.15 (−0.89,1.18)	0.2 (−0.8,1.19)	0 (−0.07,0.06)	0.01 (−0.07,0.08)	0 (−0.07,0.08)
$A_{16} = 0$	−0.82 (−2.94,1.29)	−3.07 (−5.88,−0.39)	−3.14 (−5.86,−0.5)	−0.08 (−0.28,0.12)	−0.15 (−0.45,0.15)	−0.15 (−0.45,0.15)	−0.03 (−0.09,0.03)	−0.07 (−0.15,0.02)	−0.08 (−0.17,0)
$A_{17} = 0$	−0.46 (−2.95,2.04)	−1.37 (−4.65,1.87)	−1.17 (−4.45,1.91)	−0.71 (−1.25,−0.17)	−0.59 (−1.36,0.19)	−0.57 (−1.31,0.17)	0.01 (−0.05,0.06)	0.05 (−0.02,0.12)	0.05 (−0.03,0.12)
$A_{18} = 0$	0.66 (−1.5,2.86)	0.3 (−2.65,3.15)	−0.01 (−2.73,2.67)	0.65 (0.11,1.22)	0.24 (−0.51,0.99)	0.18 (−0.54,0.92)	0.06 (0.01,0.12)	0.02 (−0.04,0.08)	0.02 (−0.04,0.08)
$A_{19} = -3$	−2.61 (−3.44,−1.78)	−4.44 (−5.58,−3.38)	−4.1 (−5.16,−3.13)	−1.46 (−1.8,−1.12)	−3.09 (−3.53,−2.66)	−2.94 (−3.36,−2.55)	−1.75 (−1.81,−1.69)	−3.06 (−3.14,−2.99)	−2.97 (−3.04,−2.9)
	8.48	11.06	10.32	1.19	0.76	0.68	0.43	0.01	0.01

Table 2: Drift parameter estimates for a two dimensional cubic model with arbitrary diffusion function. On the left is the true value of the parameter. The data used is the same as that of Table 1 sampled at the $\Delta = 0.1$ interval. In this case data is imputed to obtain the intervals $\Delta = \{0.01, 0.001\}$. The bottom of the table shows the Posterior Expected Loss in each case.

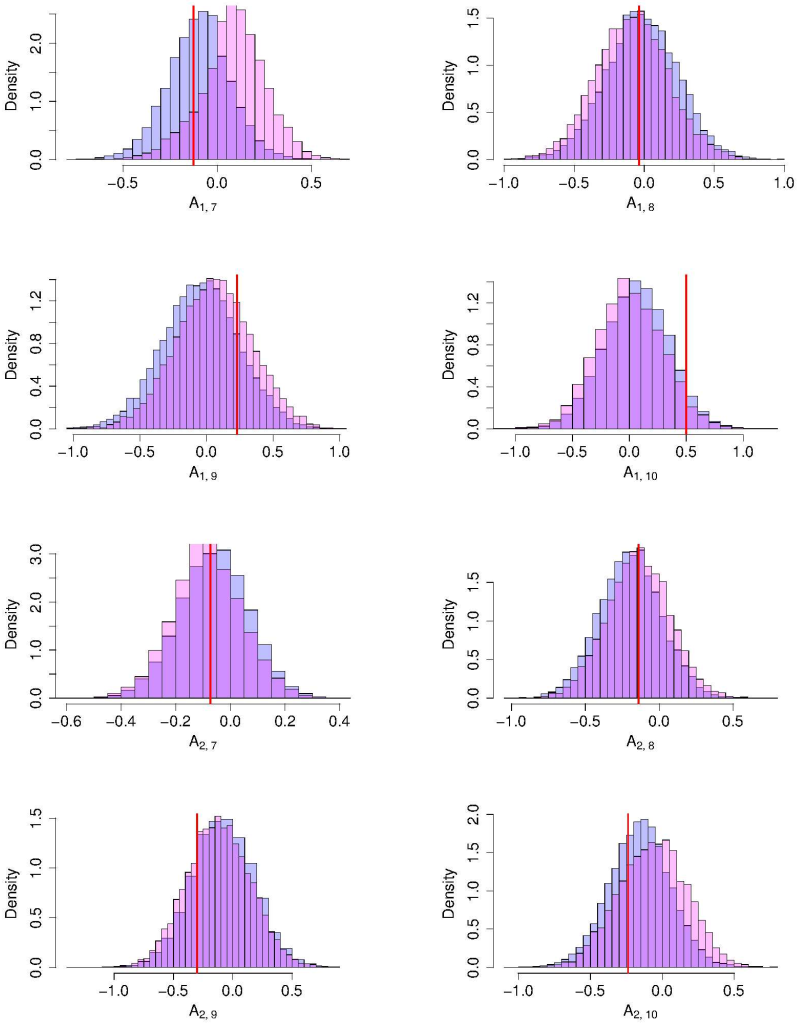


Figure 2: Marginal distributions of the cubic parameters inferred from a data set with $N = 1,000$ observations at interval $\Delta = 0.1$. The blue histogram shows the parameters that gave stable solutions to the SDE, while the red is for those that gave unstable solutions. The true values are given by the red lines.

Algorithm 3 Sample parameters along diagonal

```
for $i = 1$ to $n$ do
    $U_i$=0
    for $j = i$ to $n$ do
        $x = -\left(\sum^{j}_{\substack{k \neq i \\ k=1}} (-1)^{i+k} M_{ik} |M^{(j)}_{\{-i\},\{-k\}}|\right) / |M^{(j)}_{\{-i\},\{-i\}}|$
    end for
    if $x < U_i$ then
        $U_i = x$
    end if
    $M_{ii} \sim \mathcal{N}_{-}(\mu_i, U_i, \sigma_i^2)$
end for
```

are functions of the other parameters. These coefficients are found to be

$$a_{ij}^{(k)} = -|M^{(k)}_{/\{i,j\},/\{i,j\}}| \tag{24a}$$

$$b_{ij}^{(k)} = (-1)^{i+j} \sum_{\substack{k \neq i \\ k=1}}^{j-1} M_{jk} (-1)^{j-1+k} |M^{(k)}_{/\{i,j\},/\{j,k\}}| \tag{24b}$$

$$+(-1)^{i+j} \sum_{\substack{k \neq i \\ k=j+1}}^{N} M_{jk} (-1)^{j+k} |M^{(k)}_{/\{i,j\},/\{j,k\}}| \tag{24c}$$

$$+ \quad (-1)^{i+j} \sum_{\substack{k \neq j \\ k=1}}^{i-1} M_{ik} (-1)^{i-1+k} |M^{(k)}_{/\{i,j\},/\{i,k\}}| \tag{24d}$$

$$+(-1)^{i+j} \sum_{\substack{k \neq j \\ k=i+1}}^{N} M_{ik} (-1)^{i+k} |M^{(k)}_{/\{i,j\},/\{i,k\}}| \tag{24e}$$

$$c_{ij}^{(k)} = \sum_{\substack{k \neq j \\ k=1}}^{N} M_{ik} (-1)^{i+k} \left(\sum_{\substack{l \neq i \\ l=1}}^{k-1} M_{jl} (-1)^{j-1+l} |M^{(k)}_{/\{i,j\},/\{l,k\}}| + \right. \tag{24f}$$

$$\left. \sum_{\substack{l \neq i \\ l=k+1}}^{N} M_{jl} (-1)^{j+l} |M^{(k)}_{/\{i,j\},/\{l,k\}}| \right), \tag{24g}$$

where $|M^{(k)}_{/\{i,j\},/\{l,k\}}|$ represents the kth principal minor with rows i and j and columns l and k removed. For each component M_{ij} this quadratic form can be solved to give upper and lower bounds on the parameter. The matrix $\boldsymbol{M}$ can be cycled through updating each parameter in turn. Algorithm 4 describes the sampling of off-diagonal elements using the coefficients in Eq. (24g). Here, the notation, $\mathcal{N}_{-}^{+}(\mu, u^{-}, u^{+}, \sigma^2)$ refers to the doubly truncated normal distribution with mean μ, left truncation u^{-}, right truncation u^{+} and standard deviation σ.

To simulate from truncated normal distributions we are using the inverse Cumulative Density Function (CDF) method. One simply calculates the corresponding CDF of the lower and upper

Algorithm 4 Sample parameters off diagonal

```
for $i = 1$ to $n$ do
  for $j = i + 1$ to $n$ do
    $u^+ = \infty$
    $u^- = -\infty$
    for $k = j$ to $n$ do
      Calculate $a_{ij}^{(k)}, b_{ij}^{(k)}$ and $c_{ij}^{(k)}$ and solve $a_{ij}^{(k)}x^2 + b_{ij}^{(k)}x + c_{ij}^{(k)} = 0$.
      Set mn $= \min(x_1, x_2)$ and mx $= \max(x_1, x_2)$
    end for
    if $mx < u^+$ then
      $u^+ = mx$
    end if
    if $mn > u^-$ then
      $u^- = mn$
    end if
    $M_{ij} \sim \mathcal{N}_-^+(\mu_{ij}, u^-, u^+, \sigma_{ij}^2)$
  end for
end for
```

boundaries and then draws a uniform random variable between these numbers. Inverting the CDF gives a random variable from the Normal distribution restricted to this region.

For our problem we use the rejection sampler method proposed by (34). This method draws uncorrelated samples directly from the target density. Rejection sampling from a distribution $h(x)$ is based on a proposal distribution $g(x)$ such that $h(x) \leq Cg(x)$ holds for some constant C and all of the support of $h(x)$. For a one sided truncated Normal the exponential distribution is a good proposal. First it is translated to coincide with the truncation point, then the rate parameter is optimized in order to closely match the tail of the Normal distribution.

$$g(z; \alpha, \mu^-) = \alpha \exp(-\alpha(z - \mu^-))\mathbb{I}_{z \geq \mu^-} \tag{25}$$

The optimal value of α is calculated by maximizing the expected acceptance probability and is shown to be

$$\alpha^*(\mu^-) = \frac{\mu^- + \sqrt{(\mu^-)^2 + 4}}{2} \tag{26}$$

More details are given in (34).

We performed a numerical study to compare the standard Normal and Exponential proposals. The efficiency of proposing x from the standard normal and then accepting if $x > \mu^-$ falls to approximately 0.023 while for the optimized exponential proposal it is approximately 0.5.

For the doubly truncated Normal one uses either an exponential or uniform distribution, as a proposal, depending upon the size of the truncated region. If the following holds

$$u^+ > u^- + \frac{2\sqrt{e}}{u^- + \sqrt{(u^-)^2 + 4}} \exp(\frac{(u^-)^2 - u^- \sqrt{(u^-)^2 + 4}}{4})$$

then it can be shown that the exponential is more efficient, otherwise the uniform is better (34). Fig. 3 shows both the uniform and exponential approximations for both cases.

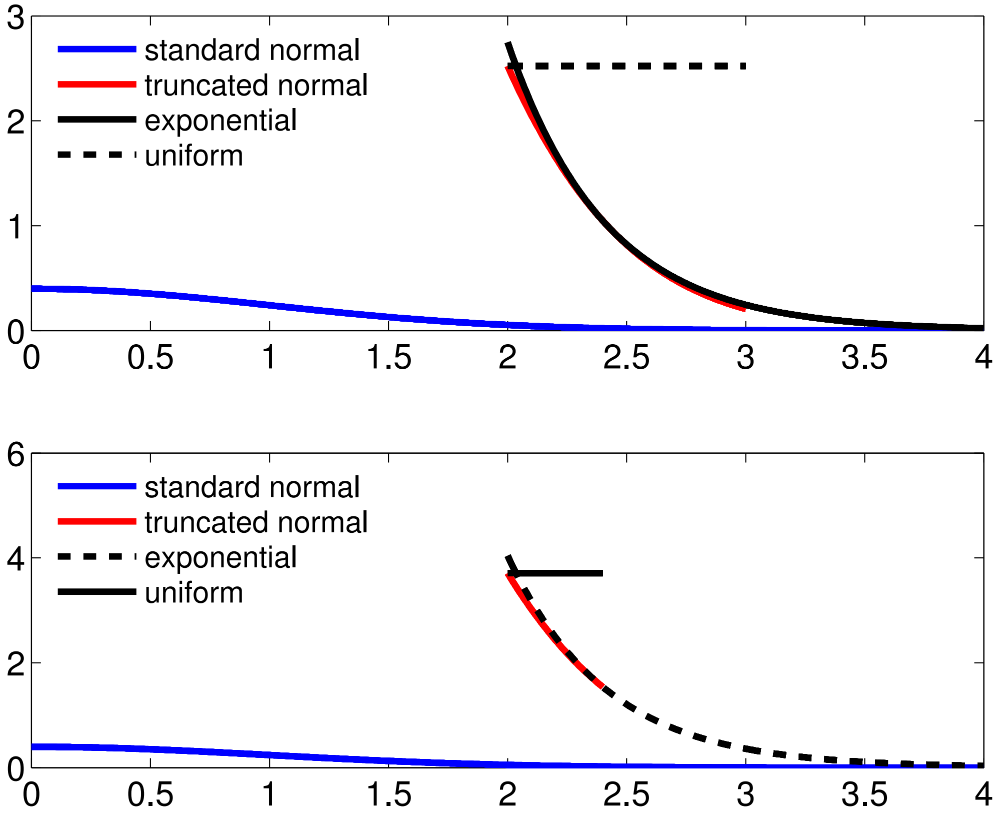


Figure 3: Doubly truncated normal distribution. The top figure has $u^- = 2$ and $u^+ = 3$ and is better approximated with the exponential distribution. The bottom figure has $u^- = 2$ and $u^+ = 2.5$ and the uniform is more efficient.

We use Algorithms 3 and 4, along with the methods of sampling truncated Normal variables, to sample the stability matrix $\boldsymbol{M}$. We tested this algorithm on a three dimensional model with $T = 100$. In this case the dimension of $\boldsymbol{M}$ is $n = D(D+1)/2 = 6$. Note that this MCMC algorithm still mixes well. Fig. 4 compares the posterior distributions estimated from the Component Wise Algorithm and the standard Gibbs sampler. Notice the large differences between the distributions in each case. Similar results (not shown) are obtained for the off diagonal parameters using Algorithm 4.

4. Results

4.1. Deterministic Double Well Potential Model

The first conceptual climate model we consider is a cubic model coupled to the chaotic Lorenz system (29). It is fully deterministic and consists of a slow variable which can be thought of as representing a climate process and three fast variables which can be thought of as representing chaotic weather fluctuations. The slow variable moves inside a double well potential and is perturbed by the chaotic Lorenz system, which acts effectively as noise when $\epsilon \to 0$. The equations are as follows

$$\frac{dx}{dt} = x - x^3 + \frac{4}{90\epsilon}y_2 \tag{27a}$$

$$\frac{dy_1}{dt} = \frac{10}{\epsilon^2}(y_2 - y_1) \tag{27b}$$

$$\frac{dy_2}{dt} = \frac{1}{\epsilon^2}(28y_1 - y_2 - y_1y_3) \tag{27c}$$

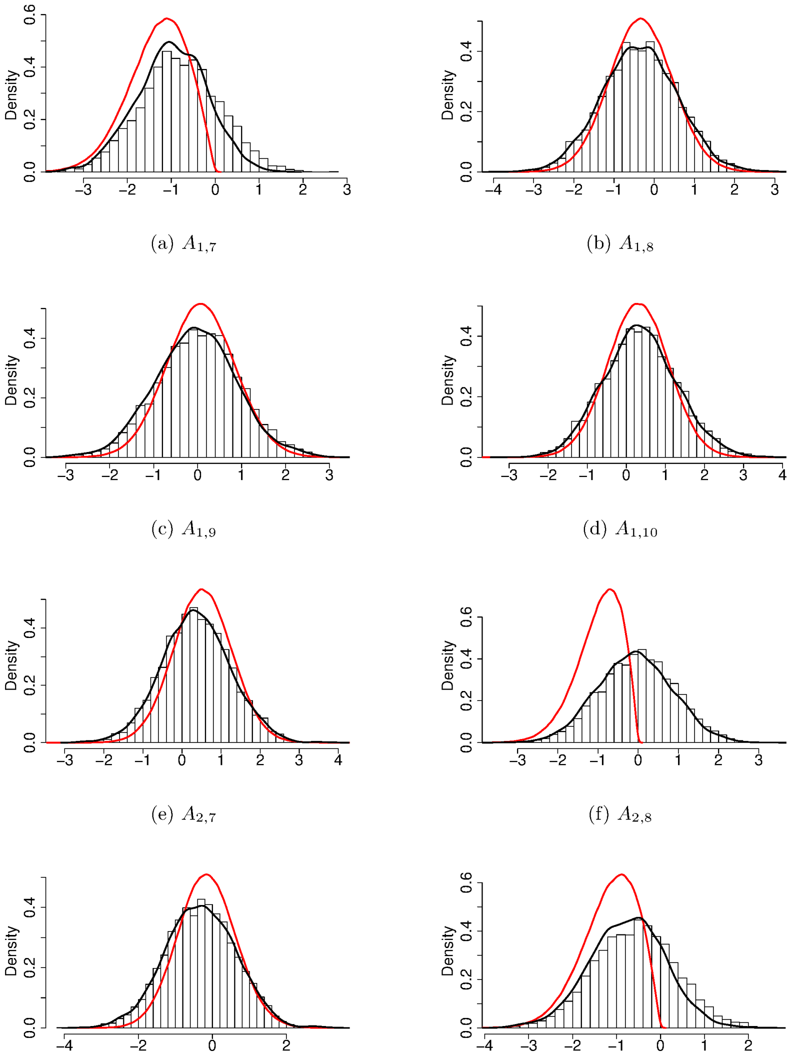
$$\frac{dy_3}{dt} = \frac{1}{\epsilon^2}(y_1y_2 - \frac{8}{3}y_3)\,. \tag{27d}$$

Sample paths are displayed in Fig. 5. We now fit a one dimensional cubic SDE to the data. We just consider the general cubic form (26)

$$dX_t = (a_1 + a_2X_t + a_3X_t^2 + a_4X_t^3)dt + \sigma dW_t \tag{28}$$

and estimate all of the parameters $\{a_1, a_2, a_3, a_4, \sigma\}$ from sparse observations of the system: again using $\Delta = 10.0$ and $N = 1000$. To update the drift parameters we use the Gibbs sampler of Section 3.2.1. The estimated posterior distributions are shown in Fig. 6. A lot of imputed data is needed before the estimates start to converge towards the values predicted by homogenization but the inference demonstrates that there is enough information in the sparse data set if the likelihood is well approximated.

Figure 7 shows the predictive skill of the one dimensional reduced model for σ estimated for various m. We use the empirical mean estimate for the parameter values. Figures 7a and 7b show that the reduced model can reproduce the double well distribution of the full model although the separation of each well is underestimated for $m = 2$ and $m = 4$ due to the larger noise. For $m \geq 8$ the model reproduces well the full models marginal distribution for x. It is not clear whether there is much difference between $m = 8$ and $m = 64$. However, observing Figures 7c and 7d we see that the autocorrelation function for the full model is much better approximated when $m = 64$. This shows that the ability of our framework to impute data is a powerful way of deriving accurate reduced order models.



(a) $A_{1,7}$

(b) $A_{1,8}$

(c) $A_{1,9}$

(d) $A_{1,10}$

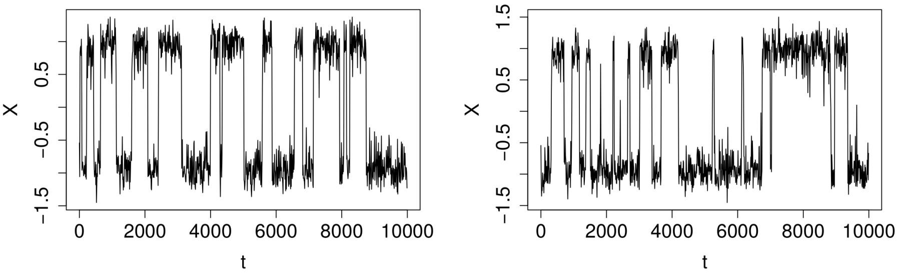
(e) $A_{2,7}$

(f) $A_{2,8}$

(g) $A_{2,9}$

(h) $A_{2,10}$

Figure 4: Estimated posterior distributions for parameters from a two dimensional model of the form Eq. 3 with N = 100 and = 0.1. The parameters, which are randomly generated, are written in the matrix notation introduced in Section 3.2.1. The histograms are the posterior distributions with uninformative prior, in red are the posterior distributions for parameters with stable SDEs and in black are the posterior distributions which include the stability matrix prior information derived in this chapter.



(a) Solution for $\epsilon = 0.1$

(b) Solution for $\epsilon = 0.01$

Figure 5: Example path of x of the chaotic Lorenz system: Eq. (27d)

4.2. Model Reduction for Triad Systems

Now we apply our model fitting procedure to a triad model with a high dimensional deterministic system with two slow, climate variables coupled to fast chaotic dynamics. The reduction strategy has two challenges: to successfully approximate the deterministic variables by a stochastic process and to be insensitive to a lack of time scale separation.

The full system is given by

$$\frac{dx_1}{dt} = \frac{b_1}{\epsilon} x_2 y_1 \tag{29a}$$

$$\frac{dx_2}{dt} = \frac{b_2}{\epsilon} x_1 y_1 \tag{29b}$$

$$\frac{dy_k}{dt} = \frac{b_3}{\epsilon} x_1 x_2 \delta_{1,k} - \mathrm{Re}\frac{ik}{2\epsilon^2} \sum_{p+q+k=0} \hat{u}_p^* \hat{u}_q^* \tag{29c}$$

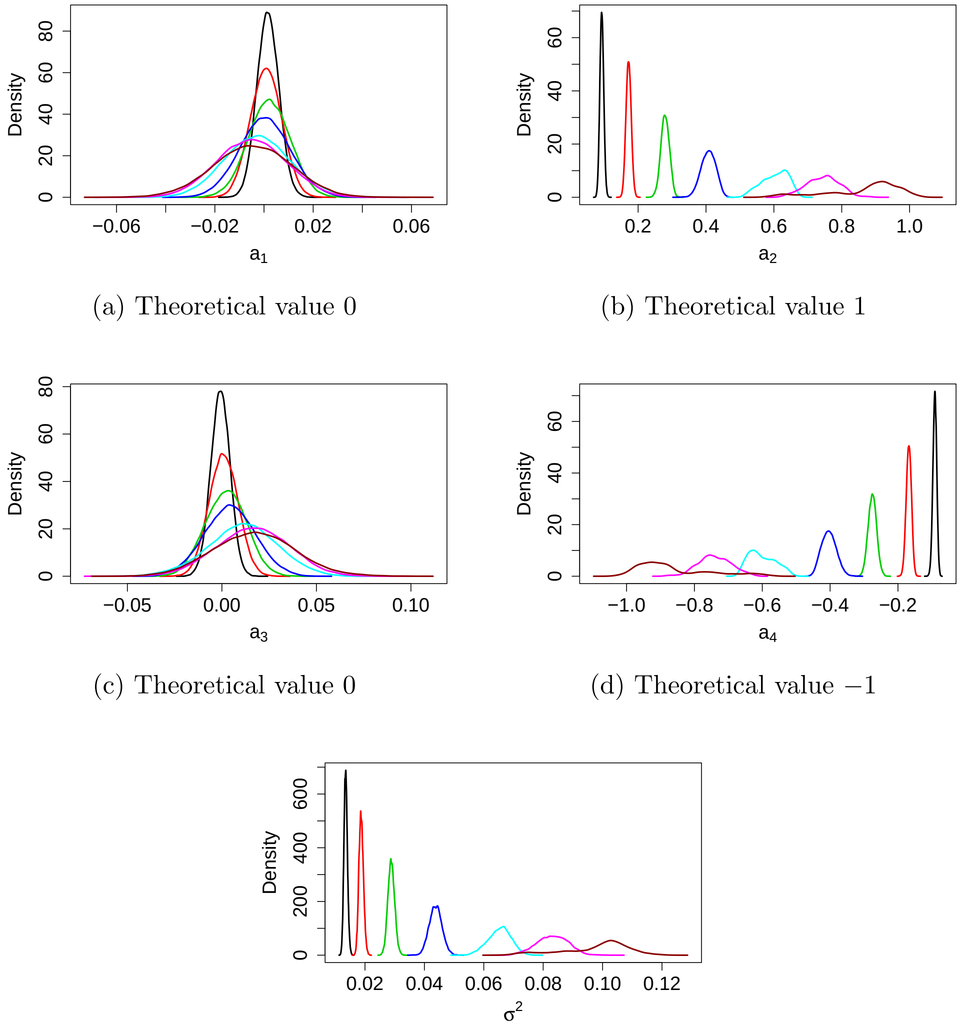
$$\frac{dz_k}{dt} = -\mathrm{Im}\frac{ik}{2\epsilon^2} \sum_{p+q+k=0} \hat{u}_p^* \hat{u}_q^*, \tag{29d}$$

where $u_k = y_k + iz_k$. This system is stable provided that the energy is conserved: $b_1 + b_2 + b_3 = 0$. We use the values $\boldsymbol{b} = \{0.9, -0.5 - 0.4\}$ (our results are insensitive over a wide range of parameter values) and we choose a cut off of $\Lambda = 50$. Sample paths are displayed in Fig. 8. We are interested in eliminating $\boldsymbol{y}$ leaving equations for just x_1 and x_2. The small parameter ϵ represents the time scales within the system. The variables $\boldsymbol{y}$ have fastest time scale of order $O(1/\epsilon^2)$ compared to $O(1/\epsilon)$ for x_1 and x_2. As $\epsilon \to 0$ we can use the method of homogenization for SDEs to eliminate the fast variables and this gives

$$dx_1(t) = \frac{b_1}{\gamma}(b_3 x_2^2(t) + \frac{\sigma^2}{2\gamma} b_2) x_1(t) dt + \frac{\sigma}{\gamma} b_1 x_2(t) dW_t \tag{30a}$$

$$dx_2(t) = \frac{b_2}{\gamma}(b_3 x_1^2(t) + \frac{\sigma^2}{2\gamma} b_1) x_2(t) dt + \frac{\sigma}{\gamma} b_2 x_1(t) dW_t\,, \tag{30b}$$

where unknown parameters σ and γ have been introduced. Here we estimate them using the Algorithms 1 and 2 from observations of the climate variables alone. For convenience we con-



(a) Theoretical value 0

(b) Theoretical value 1

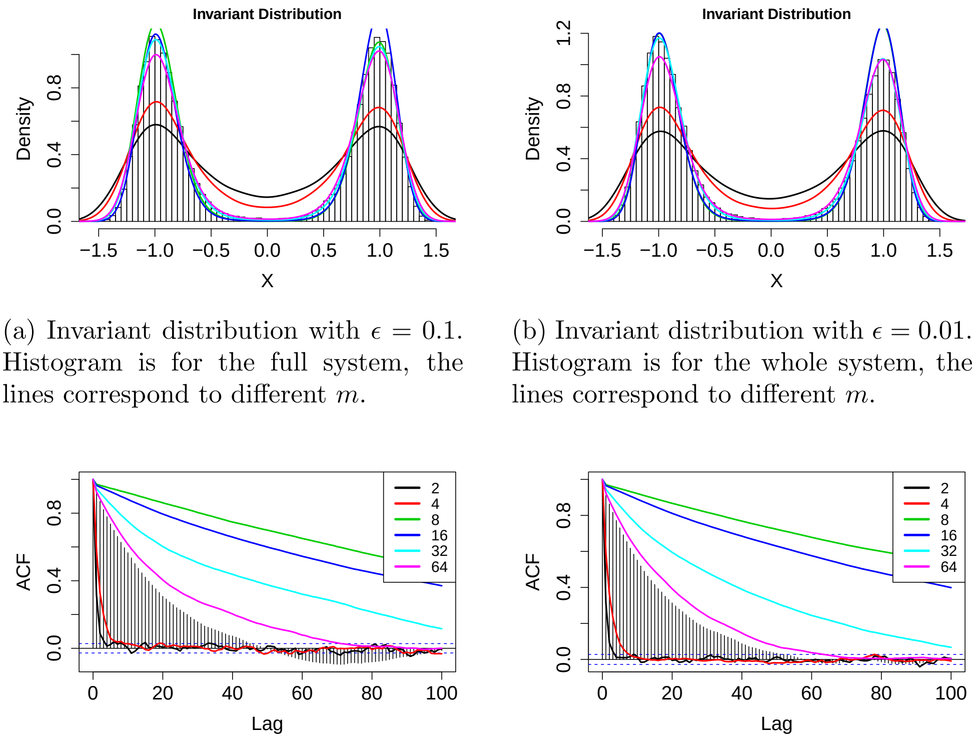
(c) Theoretical value 0

(d) Theoretical value -1

(e) Theoretical value 0.11

Figure 6: Posterior distribution estimates from MCMC output applied to a sparse data set ($\Delta = 10$). Different distributions correspond to increasing amounts of missing data. The distribution in brown, for $m = 64$, agrees with the theoretical values predicted by the homogenization procedure. In the model simulation $\epsilon = 0.01$ has been used. Black line: m=1, red line: m=2, green line: m=4, blue line: m=8, light blue line: m=16, magenta line: m=32, brown line: m=64.

sider the inference problem where the model Eq. (30) is driven by two independent Brownian



(a) Invariant distribution with $\epsilon = 0.1$. Histogram is for the full system, the lines correspond to different m.

(b) Invariant distribution with $\epsilon = 0.01$. Histogram is for the whole system, the lines correspond to different m.

(c) Autocorrelation function for $\epsilon = 0.1$. Bars are for the full system.

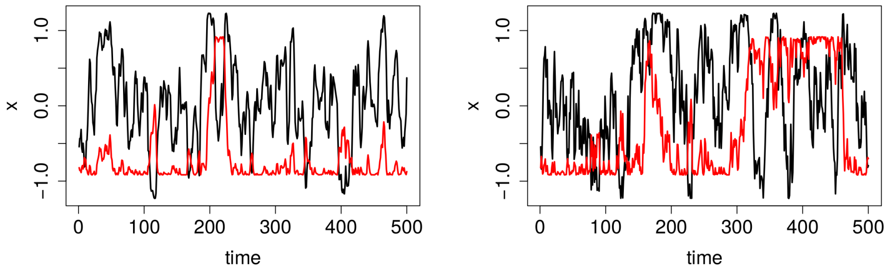
(d) Autocorrelation function for $\epsilon = 0.01$. Bars are for the full system.

Figure 7: Predictive statistics for the reduced double well model coupled to chaotic Lorenz system: Eq. (27d) for two values of ϵ. In each plot the lines correspond to the inferred one dimensional model for different m.

motions.

Posterior estimates for the case $\epsilon = 0.8$ are shown in Fig. 9. This value corresponds to a moderately small, though realistic (14), amount of time scale separation. As Fig. 9 demonstrates, increasing the number of imputed data leads to a convergence and, thus, to an improvement of the posterior estimates.

We apply the inference to a data set with total time $T = 500$ and observation interval $\Delta = 0.1$. We simulate the system for $\epsilon = \{0.1, 0.25, 0.5, 0.8, 1.0\}$. Fig. 10 shows the predictive probability densities and autocorrelation functions for Eq. 30. In each case the data is simulated from the



(a) Solution for $\epsilon = 0.1$

(b) Solution for $\epsilon = 0.01$

Figure 8: Example path of x_1 (black line) and x_2 (red line) from the triad model: Eq. (29)

full model Eq. 29, then the parameters are estimated using the reduced model Eq. 30 and this reduced model is simulated to calculate the predictive statistics. The posterior mean estimates were computed for m = 16 missing data values (it was veried that the posteriors for m = 16 and m < 16 gave consistent estimates). The reduced model Eq. 30 with empirical parameter estimates is also plotted. This model is referred to as the reduced model in Figure 10. The reduced model is able to reproduce the non-trivial shape of the PDF very well. This suggests tha reduced order models fitted from observed data can be used for extreme value studies (16).

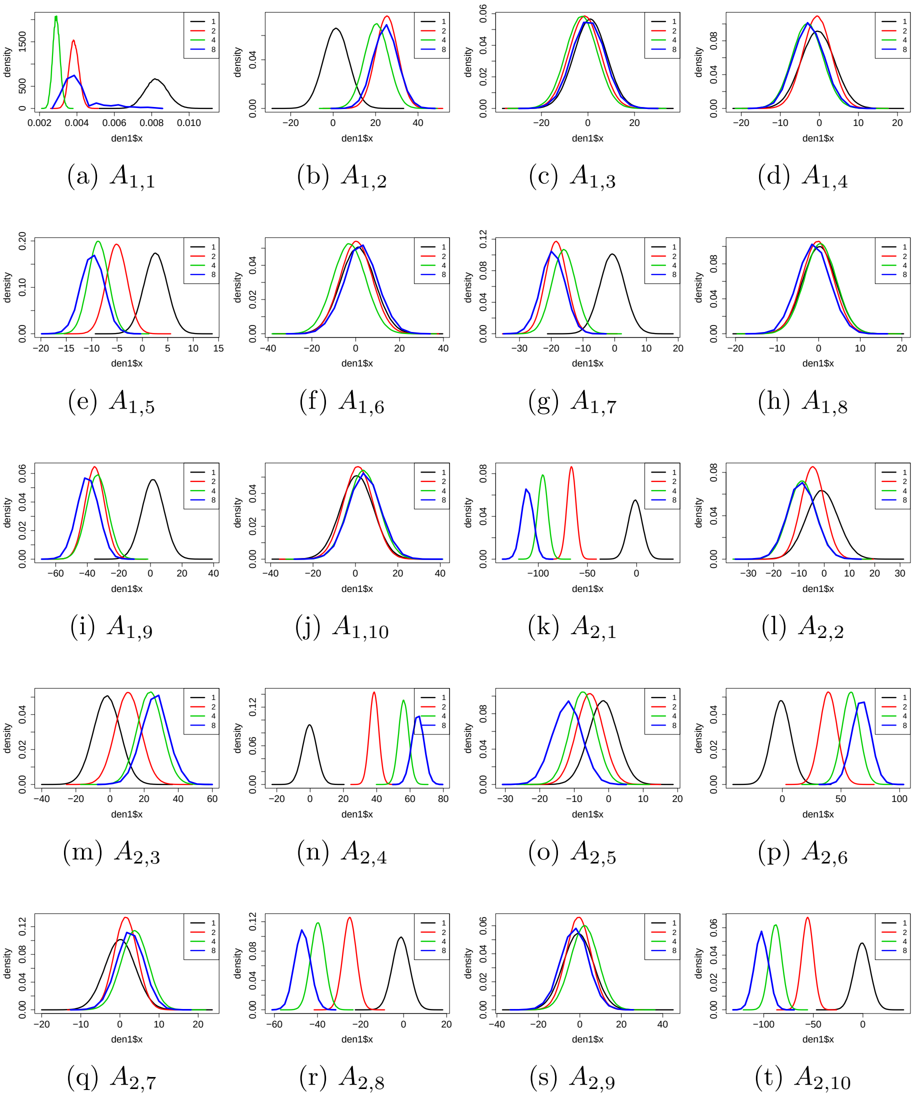
The autocorrelation functions have been collapsed onto the reduced model by rescaling the output interval of the prediction by their value of ϵ; this has been done for convenience of displaying the results. The data collapse is very good for all model simulations with the models with $\epsilon < 0.5$ being closest to the reduced model. This implies that the parameter estimates for each case are partially compensating for the changing time scale separation. This provides evidence for the potential of using reduced order modelling strategies even in systems with only moderate time scale separation.

5. Summary

Here we developed a systematic Bayesian framework for the inference of the parameter of SDEs constrained by the physics of the underlying system. The physical constraints not only constrain the parameter space but also enforce global stability of the reduced order models.

For climate models we derive a constraint based on energy conservation which ensures global stability of the effective SDE. This constraint takes the form of a negative definite matrix. We then develop a new algorithm for the sampling of negative definite matrices. We also develop a new algorithm for imputing data and show that imputing data improves the accuracy considerly of the parameter estimation. We demonstrated its power successfully on two conceptual climate models.

While we focused on climate models in this study our method is general enough that it can also be applied to other areas of fluid dynamics. Furthermore, also many other physical systems observe conservation laws and, thus, stability conditions can be derived which will be useful for parameter estimation procedures.



(a) $A_{1,1}$ (b) $A_{1,2}$ (c) $A_{1,3}$ (d) $A_{1,4}$

(e) $A_{1,5}$ (f) $A_{1,6}$ (g) $A_{1,7}$ (h) $A_{1,8}$

(i) $A_{1,9}$ (j) $A_{1,10}$ (k) $A_{2,1}$ (l) $A_{2,2}$

(m) $A_{2,3}$ (n) $A_{2,4}$ (o) $A_{2,5}$ (p) $A_{2,6}$

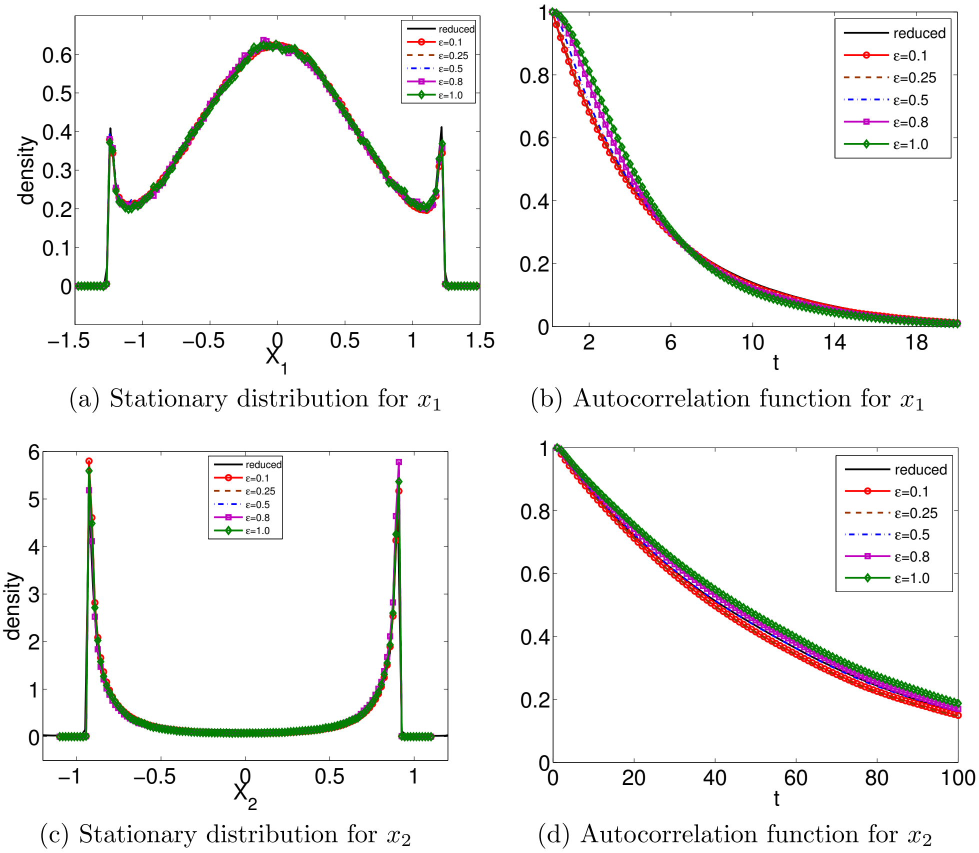
(q) $A_{2,7}$ (r) $A_{2,8}$ (s) $A_{2,9}$ (t) $A_{2,10}$

Figure 9: Posterior estimates of drift parameters for two dimensional cubic model fitted to the triad-Burgers equation for $\epsilon = 0.8$.

Acknowledgments We thank three anonymous reviewers whose comments improved earlier versions of this manuscript. This study received funding from the Natural Environment Research Council, the Engineering and Physical Sciences Research Council and the Deutsche Forschungsgemeinschaft (DFG) through the cluster of excellence CliSAP.

References

[1] Achatz, U. and G. Branstator, A two-layer model with empirical linear corrections and reduced order for studies of internal climate variability. J. Atmos. Sci., 56, (1999) 3140-3160.



(a) Stationary distribution for x_1

(b) Autocorrelation function for x_1

(c) Stationary distribution for x_2

(d) Autocorrelation function for x_2

Figure 10: Output statistics comparing the reduced model Eq. 30, with empirical models for various ϵ values with the full model.

[2] Ait-Sahalia, Y, Maximum likelihood estimation of discretely sampled diffusions: A closed-form approximation approach. Econometrica, 70, (2002) 223-262.

[3] Ait-Sahalia, Y., Closed-form likelihood expansions for multivariate diffusions. Annals Stat., 36, (2008) 906-937.

[4] Barczykern, M. and P. Kern, Representations of multidimensional linear process bridges. Rand. Oper. Stoch. Eq., 21, 159-189, (2013).

[5] Beskos, A. and G. O. Roberts, Exact simulation of diffusions. Ann. Appl. Prob., 15, 2422-2444, (2005).

[6] Beskos, A., O. Papaspiliopoulos and G. O. Roberts, A factorisation of diffusion measure and finite sample path constructions. Meth. Comp. Appl. Prob., 10, 85-104, (2008).

[7] Chib, S., Pitt, M. and Shephard, N., Likelihood based inference for diffusion driven state space models. Working Paper, Nuffield College, Oxford University, (2004).

[8] Crommelin, D. and E. Vanden-Eijnden, Reconstruction of diffusions using spectral data from timeseries. Comm. Math. Sci., 4 (2006), 651-668.

[9] Dargatz, C., Bayesian Inference for Diffusion Processes with Applications in Life Sciences. PhD thesis, Fakultät der Mathematik, Informatik und Statistik der Ludwig Maximilians Universität München, (2010).

[10] Durham, G. B. and Gallant, A. R., Numerical techniques for maximum likelihood estimation of continuous-time diffusion processes. J. Bus. Econ. Statist., 20 (2002), 297-316.

[11] Eaton, M. L., Multivariate Statistics: A Vector Space Approach. Lecture Notes–Monograph. Series, Volume 53.

[12] Eraker, B., MCMC analysis of diffusion models with application to finance. J. Bus. Econ. Statist., 19 (2001), 177-191.

[13] Elerian, O. S., Simulation estimation of continuous-time models with applications to finance. PhD thesis, Nuffield College, Oxford (1999).

[14] Franzke, C., A. J. Majda and E. Vanden-Eijnden, Low-Order Stochastic Mode Reduction for a Realistic Barotropic Model Climate. J. Atmos. Sci., 62 (2005), 1722-1745.

[15] Franzke, C., and A. J. Majda, Low-Order Stochastic Mode Reduction for a Prototype Atmospheric GCM. J. Atmos. Sci., 63 (2006), 457-479.

[16] Franzke, C., Predictability of Extreme Events in a Nonlinear Stochastic-Dynamical Model. Phys. Rev. E, 85 (2012), DOI: 10.1103/PhysRevE.85.031134

[17] Friedrich, R., S. Siegert, J. Peinke, St. Lück, M. Siefert, M. Lindemann, J. Raethjen, G. Deuschl and G. Pfister, Extracting model equations from experimental data. Phys. Lett. A, 271 (2000), 217-222.

[18] Golightly, A. and D. J. Wilkinson, Bayesian inference for nonlinear multivariate diffusion models observed with error. Comp. Stat. Data Anal., 52 (2008), 1674-1693.

[19] Harlim, J., A. Mahdi and A. J. Majda, An ensemble Kalman filter for statistical estimation of physics constrained nonlinear regression models. J. Comp. Phys., 257, (2014), 782-812.

[20] Heinz, S., Statiscal mechanics of turbulent flows. Springer Verlag Berlin, (2014), 240pp.

[21] Horenko, I., E. Dittmer and C. Schütte, Reduced stochastic models for complex molecular systems. SIAM Comp. Vis. Sci., 9, (2005), 789-102.

[22] Kondrashov, D., S. Kravtsov and M. Ghil, Empirical mode reduction in a model of extratropical low-frequency variability. J. Atmos. Sci., 63 (2006), 1859-1877.

[23] Majda, A. J., I. Timofeyev and E. Vanden-Eijnden, Models for stochastic climate prediction. Proc. Nat. Acad. Sci. USA, 96 (1999), 14687-14691.

[24] Majda, A. J., I. Timofeyev and E. Vanden-Eijnden, A mathematical framework for stochastic climate models. Commun. Pure Appl. Math., 54 (2001), 891-974.

[25] Majda, A. J., C. Franzke and B. Khouider, An applied mathematics perspective on stochastic modelling for climate. Phil. Trans. R. Soc. A, 366 (2008), 2429-2455.

[26] Majda, A. J., C. Franzke and D. Crommelin, Normal forms for reduced stochastic climate models. Proc. Nat. Acad. Sci. USA, 106, (2009) 3649-3653. doi: 10.1073/pnas.0900173106

[27] Majda, A. J. and Y. Yuan, Fundamental limitations of ad hoc linear and quadratic multi-level regression models for physical systems. Disc. Con. Dyn. Sys., 17, (2012) 1333-1363.

[28] Majda, A. J. and J. Harlim, Physics constrained nonlinear regression models for time series. Nonlinearity, 26, (2013) 201-217.

[29] Mitchell, L. and Gottwald, G. A., Data Assimilation in Slow-Fast Systems Using Homogenized Climate Models. J. Atmos. Sci., 69, (2012) 1359-1377.

[30] Ozaki, T., A bridge between nonlinear time series models and nonlinear stochastic dynamic systems. A local linearization approach. Stat. Sinica, 2, (1992) 113-135.

[31] Pavliotis, G. A. and A. M. Stuart, Multiscale methods, Springer Verlag, (2008) 310pp.

[32] Peavoy, D., Methods of likelihood based inference for constructing stochastic climate models, PhD thesis, University of Warwick, (2013) 232pp.

[33] Pedersen, A. R., A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. Scand. J. Statist., 22, (1995) 55-71.

[34] Robert, C., Simulation of truncated normal variables. Stat. Comp., 5, (1995) 121-125.

[35] Roberts, G. O., A. Gelman and W. R. Gilks, Weak convergence and optimal scaling of random walk Metropolis algorithms. Ann. App. Prob., 7, (1997) 110-120.

[36] Roberts, G. O. and Stramer, O., On inference for partially observed nonlinear diffusion models using the Metropolis-Hastings algorithm. Biometrika, 88, (2001) 603-621.

[37] Selten, F. M., An efficient description of the dynamics of barotropic flow. J. Atmos. Sci., 52, (1995) 915-936.

[38] Shephard, N. and Pitt, M. K., Likelihood analysis of non-Gaussian measurement time series. Biometrika, 84, (1997) 653-667.

[39] Shoji, I. and T. Ozaki, Estimation for nonlinear stochastic differential equations by a local linearization method. Stoch. Anal. Appl., 16, (1998) 733-752.

[40] Siegert, S., R. Friedrich and J. Peinke, Analysis of data sets of stochastic systems. Phys. Lett. A, 243 (1998), 275-280.

[41] Yuan, Y. and A. J. Majda, Invariant measures and asymptodic Gaussian bounds for normal forms of stochastic climate models. Chin. Ann. Math., 32 (2011), 343-368.