Data-driven stochastic representations of unresolved features in multiscale models

Verheul, N.; Crommelin, D.

DOI
10.4310/CMS.2016.v14.n5.a2

Publication date
2016

Document Version
Final published version

Published in
Communications in Mathematical Sciences

COMMUN. MATH. SCI.
Vol. 14, No. 5, pp. 1213–1236

DATA-DRIVEN STOCHASTIC REPRESENTATIONS OF UNRESOLVED FEATURES IN MULTISCALE MODELS*

NICK VERHEUL† AND DAAN CROMMELIN‡

Abstract. In this study, we investigate how to use sample data, generated by a fully resolved multiscale model, to construct stochastic representations of unresolved scales in reduced models. We explore three methods to model these stochastic representations. They employ empirical distributions, conditional Markov chains, and conditioned Ornstein–Uhlenbeck processes, respectively. The Kac–Zwanzig heat bath model is used as a prototype model to illustrate the methods. We demonstrate that all tested strategies reproduce the dynamics of the resolved model variables accurately. Furthermore, we show that the computational cost of the reduced model is several orders of magnitude lower than that of the fully resolved model.

Key words. Multiscale modeling, stochastic model reduction, Kac–Zwanzig heat bath model.

AMS subject classifications. 65C20, 37M05, 60H35, 60H10.

1. Introduction

1.1. Background and motivation. Multiscale modeling is an active research topic in such fields as biomedical engineering, materials science and climate modeling. The common property of multiscale problems is the occurrence of a wide range of spatial and/or temporal scales, often resulting in an inability of numerical simulations to accurately resolve the small and/or fast scales. However, processes at these scales can be instrumental in driving the large scale processes, hence they must be represented in a simplified yet accurate manner in numerical models.

The motivation for this study comes primarily from atmosphere–ocean science, where the problem of formulating suitable representations of unresolved processes is well-known. In the field of atmosphere–ocean modeling, such representations are known under the name *parameterizations*. In this field, early developments on multiscale problems used deterministic methods to represent the effect of unresolved processes. However, although deterministic methods can reproduce the mean effect of the unresolved processes conditioned on the resolved variables, they lack the ability to reproduce the fluctuations around this mean. Recent work has focused on overcoming this limitation by using stochastic methods to model this noise-like behavior, particularly in atmospheric context [8, 9, 11, 14, 18]. Notable examples for the present study include [3] and [4], which propose data-inferred conditional Markov chains to represent atmospheric convection in coarse climate models. Recently, stochastic parameterizations have also started to receive attention in oceanic research, e.g. [1, 2] and [15], which investigate stochastic eddy-forcing in ocean currents.

In this study we investigate data-driven stochastic methods to drive *reduced* multiscale models. In atmosphere–ocean modeling, there are many scales but no strong scale separation (or scale gap), so that techniques that rely on such a scale gap to achieve

*Received: December 19, 2014; accepted (in revised form): August 16, 2015. Communicated by Shi Jin.

†Centrum Wiskunde and Informatica (CWI), Science Park 123, 1098XG Amsterdam, The Netherlands (nick.verheul@cwi.nl).

‡CWI, Science Park 123, 1098XG Amsterdam, The Netherlands. And KdV Institute for Mathematics, University of Amsterdam, Science Park 105, 1098XG Amsterdam, The Netherlands (daan.crommelin@cwi.nl).

computational efficiency gains (e.g. averaging, equation-free modeling [10], heterogeneous multiscale methods [6]) are less attractive. A data-driven approach can be an interesting alternative in such cases. The idea of such an approach is to infer a suitable stochastic process from data (time series) of the feedback from the small/fast scales, and to couple this process to a reduced model for the large/slow scales. The statistical inference step is performed offline, i.e. the stochastic process for the unresolved scales is precomputed. Thus, it can be considered a "sequential coupling" method [6]. As we will demonstrate, the computational gain of this data-driven methodology can be very substantial.

We emphasize that the methodology studied here is different from inferring a stochastic process for the large scale dynamics itself. Rather, it is aimed at situations where an available but incomplete model for the large scale dynamics needs to be augmented with a model for small scale feedbacks (as is the case in e.g. atmosphere–ocean modeling). In general, a suitable stochastic model for the small scale feedbacks must be dependent (conditioned) on the state of the large scale degrees of freedom. The statistical inference step for such a conditioned stochastic process is not straightforward. We approach this issue by considering the large scale state as a covariate for the stochastic process that needs to be inferred.

The data-driven methodology studied in this paper builds on the work presented in [3]. There, finite-state Markov chains were used to model feedback from unresolved scales in the context of the Lorenz '96 model. This conditional Markov chain approach gave good results but involved the estimation of many parameters. Furthermore, in [3] no experiments were performed with different sets of conditioning variables (or covariates). In the current study we explore methods that require far less parameters to be estimated (or even none at all). For completeness, a method that stays close to [3] is included in this exploration. We also investigate the effect that varying the set of conditioning variables has on the resulting reduced model.

In the remainder of the introduction we formally pose the discussed problem and the questions this work attempts to answer. Section 2 describes the prototype multiscale model and details on its numerical implementation. Section 3 presents the three different strategies used to fit the stochastic process to the sample data: the empirical, conditional Markov chain and Ornstein–Uhlenbeck approaches, respectively. Lastly, the results and their implications for future work are discussed in Section 4.

1.2. Problem description. Given a stationary time series $\boldsymbol{X} = (\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_M)$, for $\boldsymbol{x}_i \in \mathbb{R}^d$, we wish to formulate a model such that when we integrate this model numerically, we generate a time series $\tilde{\boldsymbol{X}} = (\tilde{\boldsymbol{x}}_0, \tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_N)$, for $\tilde{\boldsymbol{x}}_i \in \mathbb{R}^d$, whose statistics accurately resemble those of $\boldsymbol{X}$. Throughout this paper we compare given data sets, where variables are denoted normally (e.g. $\boldsymbol{x}$), with data sets, denoted with a tilde (e.g. $\tilde{\boldsymbol{x}}$), generated by reduced models.

For the stochastic approach discussed here we assume that the given sample data consists of both $\boldsymbol{X}$ and $\boldsymbol{R}$, where $\boldsymbol{R}$ represents small-scale features. As an example, one can think of fluid flow, with $\boldsymbol{X}$ and $\boldsymbol{R}$ time series of the resolved-scale flow and the subgrid-scale stress term, respectively. Let $\tilde{\boldsymbol{X}}$ be generated by a reduced model g together with a stochastic process $\tilde{\boldsymbol{R}} = (\tilde{\boldsymbol{r}}_0, \tilde{\boldsymbol{r}}_1, \ldots, \tilde{\boldsymbol{r}}_N)$, for $\tilde{\boldsymbol{r}}_i \in \mathbb{R}^d$, that is fitted to $\boldsymbol{R}$. This construction describes the class of systems:

$$\dot{\tilde{\boldsymbol{x}}} = g(\tilde{\boldsymbol{x}}) + \tilde{\boldsymbol{r}}, \; \dot{\tilde{\boldsymbol{r}}} = h(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{r}}), \tag{1.1}$$

where $\dot{\tilde{\boldsymbol{x}}}$ denotes the temporal derivative of $\tilde{\boldsymbol{x}}$ (and analogously for $\dot{\tilde{\boldsymbol{r}}}$). This class of systems finds practical applications in, e.g, modeling the eddy forcing term with $\tilde{\boldsymbol{r}}$ in

ocean flow models [1], and was the inspiration for this work.

Note that we assume analytic solutions to the discussed problem to be unknown. Therefore, we will make use of numerical integration schemes. Let us introduce the following notations: $t_i = i\Delta t$, $\boldsymbol{x}_i = \boldsymbol{x}(t_i)$ denotes the $(i+1)$th entry in the time series $\boldsymbol{X}$, and $\Delta \boldsymbol{x}_i = \boldsymbol{x}_{i+1} - \boldsymbol{x}_i$.

Although we have no rigorous proof, we expect the statistics of $\boldsymbol{X}$ to be accurately emulated by $\tilde{\boldsymbol{X}}$ if it were possible to sample $\tilde{\boldsymbol{r}}_{i+1} = \tilde{\boldsymbol{r}}(t_{i+1})$ from the conditional distribution of $r_{i+1} | (\boldsymbol{x}_i = \tilde{\boldsymbol{x}}_i, \ldots, \boldsymbol{x}_0 = \tilde{\boldsymbol{x}}_0, \boldsymbol{r}_i = \tilde{\boldsymbol{r}}_i, \ldots, \boldsymbol{r}_0 = \tilde{\boldsymbol{r}}_0)$. In general, however, such distributions are not known exactly, and the size of sample data needed to accurately approximate conditional distributions increases drastically with the number of conditions. Therefore, we investigate how well the statistics of $\tilde{\boldsymbol{X}}$ approximate those of $\boldsymbol{X}$ when conditioning $\tilde{\boldsymbol{r}}_{i+1}$ on a selection of past values of $\boldsymbol{x}$ and $\boldsymbol{r}$. The approximation quality of $\tilde{\boldsymbol{X}}$ is measured by the degree to which specific sample moments and autocorrelations of $\boldsymbol{X}$ are captured by $\tilde{\boldsymbol{X}}$.

Formally, let $\tilde{\boldsymbol{r}}_{i+1}$ be sampled from the distribution of $r_{i+1} | (\boldsymbol{x}_i = \tilde{\boldsymbol{x}}_i, \ldots, \boldsymbol{x}_{i-i'} = \tilde{\boldsymbol{x}}_{i-i'}, \boldsymbol{r}_i = \tilde{\boldsymbol{r}}_i, \ldots, \boldsymbol{r}_{i-i''} = \tilde{\boldsymbol{r}}_{i-i''})$, with $0 \leqslant i', i'' \leqslant i$, and consider the following questions:

- Let the sample mean and standard deviation of $\boldsymbol{X}$ be denoted by $\gamma_1(\boldsymbol{X}) = \mathbf{E}(\boldsymbol{x}_i)$ and $\gamma_2(\boldsymbol{X}) = (\mathbf{E}(\boldsymbol{x}_i^2) - \mathbf{E}(\boldsymbol{x}_i)^2)^{1/2}$, respectively (with $\mathbf{E}$ denoting expectation). Let the sth sample moment of $\boldsymbol{X}$ (with $s \geqslant 3$) be given by

$$\gamma_s(\boldsymbol{X}) = \mathbf{E}\left[(\boldsymbol{x}_i - \mathbf{E}(\boldsymbol{x}_i))^s\right] (\mathrm{Var}(\boldsymbol{x}_i))^{-s/2}.$$

  Let $\epsilon(\gamma_s) := \gamma_s(\boldsymbol{X}) - \gamma_s(\tilde{\boldsymbol{X}})$ be the error of the sth sample moment as reproduced by $\tilde{\boldsymbol{X}}$, and let S be the maximum moment one aims to reproduce. How does $\epsilon(\gamma_s)$ depend on the number of past values of $\boldsymbol{x}$ and $\boldsymbol{r}$ conditioning $\boldsymbol{r}_{i+1}$, i.e. how does $\epsilon(\gamma_s)$ depend on i' and i''? Particularly, let E denote a maximum error one is willing to permit, for what i' and i'' does $\epsilon(\gamma_s) \leqslant E$ hold for $1 \leqslant s \leqslant S$?
- Let the autocorrelation function of $\boldsymbol{X}$ with lag l be given by

$$\mathrm{ACF}_l(\boldsymbol{X}) = \mathbf{E}\left[(\boldsymbol{x}_i - \mathbf{E}(\boldsymbol{x}_i))(\boldsymbol{x}_{i+l} - \mathbf{E}(\boldsymbol{x}_i))\right] (\mathrm{Var}(\boldsymbol{x}_i))^{-1}.$$

  Let $\epsilon(\mathrm{ACF}_l) := \mathrm{ACF}_l(\boldsymbol{X}) - \mathrm{ACF}_l(\tilde{\boldsymbol{X}})$ be the error of the autocorrelation with lag l as reproduced by $\tilde{\boldsymbol{X}}$, and let L be the maximum correlation lag time one aims to reproduce. How does $\epsilon(\mathrm{ACF}_l)$ depend on i' and i''? Particularly, let E' denote a maximum error one is willing to permit, for what i' and i'' does $\epsilon(\mathrm{ACF}_l) \leqslant E'$ hold for $0 \leqslant l \leqslant L$?

Rather than dealing with the technical intricacies and complications of testing methodologies directly on highly complex multiscale models, we elect to test our ideas on the simpler and more accessible Kac–Zwanzig *heat bath model* [7, 19]. This model, described below, also belongs to the class of systems in (1.1).

Assume a resolved heat bath model's sample data, $(\boldsymbol{X}, \boldsymbol{R}) = (\boldsymbol{Q}, \boldsymbol{P}, \boldsymbol{R})$, where $\boldsymbol{Q} = (q_0, q_1, \ldots, q_M)$, $\boldsymbol{P} = (p_0, p_1, \ldots, p_M)$, and $\boldsymbol{R} = (r_0, r_1, \ldots, r_M)$, for $q_i, p_i, r_i \in \mathbb{R}$, is given. The question we attempt to answer here is: "How can we fit a stochastic process $\tilde{\boldsymbol{R}}$ to $\boldsymbol{R}$ in such a way that the reduced model variables' time series, $\tilde{\boldsymbol{Q}}$ and $\tilde{\boldsymbol{P}}$, reproduce the statistics of $\boldsymbol{Q}$ and $\boldsymbol{P}$, respectively?" With respect to this heat bath model, a thorough theoretical analysis of the questions asked in this section eludes us. Therefore, we approach these questions from a numerical perspective.

2. Kac–Zwanzig heat bath: a prototype model

2.1. Model description. In the heat bath model, one considers the temporal evolution of a distinguished particle, moving in a potential V and coupled to J heat bath particles. The distinguished particle has unit mass, position q, and momentum p. We use the set-up from [16], with a double-well potential $V(q) = 1/4(q^2-1)^2$ and linear coupling of the heat bath particles to the distinguished particle. The heat bath particles are oscillators, each with their own momentum u_j, position v_j, mass χ_j and stiffness ξ_j, with $1 \leqslant j \leqslant J$. Following [16], let us define the oscillators' natural frequency through $\omega_j^2 = \xi_j/\chi_j$, and choose the oscillator mass $\chi_j = G^2/j^2$ and stiffness $\xi_j = G^2$. The considered heat bath model's Hamiltonian system is then given by the following ordinary differential equations (ODEs):

$$\dot{q} = p, \quad \dot{p} = -V'(q) + G^2(r - Jq), \quad \dot{u_j} = v_j, \quad \dot{v_j} = -j^2(u_j - q), \tag{2.1}$$

where $V'(q) = dV(q)/dq$ and $r(t) := \sum_{j=1}^{J} u_j(t)$. While these ODEs can be solved numerically, the computational cost of evolving p and, more importantly, every u_j and v_j over time will significantly slow down any numerical solver. Therefore, to decrease the required computational work, we introduce a stochastic process $\tilde{\boldsymbol{R}}$ that approximates the dynamical effect of $\boldsymbol{R}$. Writing r_m for $\sum_j u_j(t_m)$, we have

$$\boldsymbol{R} = (r_0, r_1, \ldots, r_M).$$

By using $\tilde{\boldsymbol{R}}$ instead of $\boldsymbol{R}$, the heat bath particles (i.e., u_j and v_j) no longer need to be evolved, thus reducing the full system in (2.1) to:

$$\dot{\tilde{q}} = \tilde{p}, \quad \dot{\tilde{p}} = -V'(\tilde{q}) + G^2(\tilde{r} - J\tilde{q}), \quad \dot{\tilde{r}} = h(\tilde{q}, \tilde{p}, \tilde{r}), \tag{2.2}$$

where the function h that evolves $\tilde{r}$ over time is yet to be defined.

As mentioned in 1.2, this construction is meant to provide our strategies with a test bed that naturally extends to geophysical fluid flow models. With this in mind, let us motivate our choice for the heat bath model. First, the heat bath particles span a great variety of time scales without a scale gap (because the natural frequencies range from $O(1)$ to $O(J)$), similar to the range of time scales in ocean flow models (as mentioned in 1.1). Also, the reduced heat bath (2.2) and reduced ocean flow models [1] belong to the same class of systems (1.1), in the sense that the stochastic term $\tilde{r}$ enters in an additive fashion (i.e. $\tilde{r}$ is added linearly to the ODE for $\tilde{\boldsymbol{x}}$, there is no multiplication with a function of $\tilde{\boldsymbol{x}}$). These reasons, together with its technical simplicity, make the heat bath model a suitable choice for our experiments. We remark that we do not attempt to preserve the Hamiltonian structure or the conserved quantities of (2.1) in the reduced model, as this is less relevant for applications in geophysical fluid flow. Furthermore, we do not consider the limit $J \to \infty$, as is done in e.g. [16], rather we keep J fixed at a finite value.

2.2. Numerical integration schemes. System (2.1) is integrated in time using the symplectic Euler method, which correctly resolves the distinguished particle's motion under the condition $\omega_j \Delta t = O(1)$ [16]. Table 2.1 shows all model parameter settings used for the simulations in this paper. The discretized integration scheme for (2.1) is the following:

$$\begin{aligned} p_{i+1} &= p_i - \Delta t\ V'(q_i) + \Delta t\ G^2(r_i - Jq_i), & v_{i+1,j} &= v_{i,j} - \Delta t\ j^2(u_{i,j} - q_i), \\ q_{i+1} &= q_i + \Delta t\ p_{i+1}, & u_{i+1,j} &= u_{i,j} + \Delta t\ v_{i+1,j}. \end{aligned}$$

Let $\mathcal{N}(x,y^2)$ denote a normal distribution with mean x and variance y^2; the harmonic oscillators are initialized by $v_j(0)=0$ and $u_j(0)\sim\mathcal{N}(0,1/(\beta k_j))$. The distinguished particle is initialized at $q_0=1$ and $p_0=0$.

Because of the chosen values for ω_j and the condition $\omega_j\Delta t=O(1)$, one sees that $J\Delta t=O(1)$ must also hold. This means that Δt must decrease as J increases for the symplectic integration scheme to properly resolve all the heat bath particle's scales. Since u_j and v_j are not evolved in the reduced model, the integration time step of a reduced simulation can generally be chosen to be much larger. Therefore, we make a distinction between Δt and $\Delta\tau$ to refer to the time steps of the resolved and reduced model, respectively. Furthermore, the resolved time series is stored with a sampling interval δt ($\geqslant\Delta t$), see Table 2.1. Recall from Section 1.2 that, throughout this paper, we use the notation $\tilde{q}$ to refer to a variable in the reduced model that is the counterpart of the variable q in the fully resolved model. Discretizing (2.2) results in the following integration scheme for the reduced model:

$$\begin{aligned}\tilde{p}_{i+1}&=\tilde{p}_i-\Delta\tau\ V'(\tilde{q}_i)+\Delta\tau\ G^2(\tilde{r}_i-J\tilde{q}_i),\\ \tilde{q}_{i+1}&=\tilde{q}_i+\Delta\tau\ \tilde{p}_{i+1},\\ \tilde{r}_{i+1}&=\tilde{r}_i+\Delta\tau\ h(\tilde{q}_i,\tilde{p}_i,\tilde{r}_i),\end{aligned}\tag{2.3}$$

where the initial conditions are chosen to be $\tilde{p}_0=p_0$, $\tilde{q}_0=q_0$ and $\tilde{r}_0=r_0$. The function h in (2.3) is not known analytically, but will be inferred from the data $(\boldsymbol{Q},\boldsymbol{P},\boldsymbol{R})$. The different stochastic methods proposed here all aim to model $\tilde{\boldsymbol{R}}$ in such a way that $\tilde{\boldsymbol{Q}}$ and $\tilde{\boldsymbol{P}}$ together with $\tilde{\boldsymbol{R}}$ reproduce the statistics of $\boldsymbol{Q}$ and $\boldsymbol{P}$. In the next section we discuss the binning procedure used in our methods.

	Parameter	Resolved model	Reduced model
G^2	mass and stiffness scaling	1	1
β	inverse temperature	10^{-4}	–
J	number of harmonic oscillators	10^2	–
M	number of sample points	10^7	10^7
δt	sampling interval	10^{-2}	10^{-2}
Δt	integration time step resolved model	10^{-4}	–
$\Delta\tau$	integration time step reduced model	–	10^{-2}
N_B	number of bins per continuous conditioning variable	10	10

TABLE 2.1. *Heat bath model parameters.*

2.3. Approximating conditional distributions by binning. In the reduced model (2.3), $\boldsymbol{R}$ is approximated with the random process $\tilde{\boldsymbol{R}}$. The strategies discussed in this paper sample $\tilde{r}$ from the distribution of r conditioned on a set of resolved model variables $c:=c(q,p,r)$:

$$\tilde{r}_{i+1}\sim r_{i+1}\,|\,(c_i=\tilde{c}_i).\tag{2.4}$$

A simple example is $c_i=\{r_i\}$; in this case $\tilde{r}_{i+1}$ is a time-correlated stochastic process. In this work, we consider different methods of approximating the distribution $r_{i+1}\,|\,(c_i=\tilde{c}_i)$, or $r_{i+1}\,|\,c_i$ for short, because the exact distribution is usually unknown. The majority of these methods approximate this distribution using a binning procedure, as explained further below.

Let us consider a set of conditioning variables c_i with cardinality $C+D$, where C and D are the number of *continuous* and *discrete* conditioning variables, respectively. The discrete variables only apply to the CMC approach, and are discussed in Section 3.2 (in other sections $D=0$ holds). The range between the minimum and maximum of each continuous conditioning variable is then independently partitioned in N_B equidistant intervals. This partitioning results in C-dimensional disjoint bins α_b, where $1 \leqslant b \leqslant B := (N_B)^C$. Each of these bins describes a set of r_{i+1}-values ρ_b, also with $1 \leqslant b \leqslant B$. This procedure is illustrated in Figure 2.1 for the case $c_i = \{q_i\}$ in (2.4). This figure shows that through discretizing the q_i-domain, one finds a mapping from intervals over q_i to sets of r_{i+1}-values.

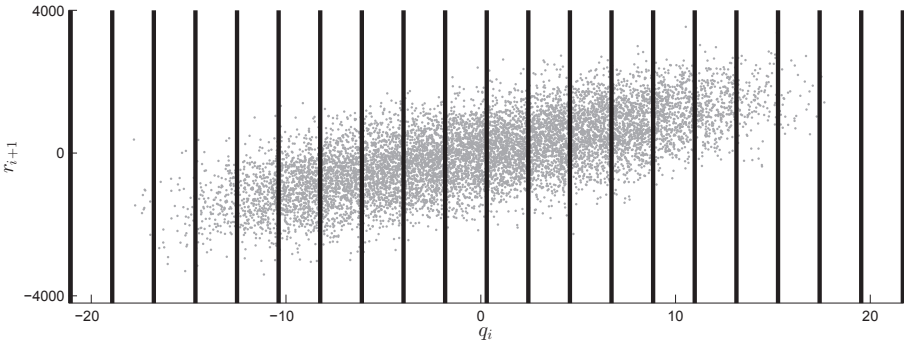

FIG. 2.1. *An equidistant partitioning of the range of q in 20 bins.*

The major advantage of the equidistant binning strategy is its simplicity in both concept and implementation. A caveat is that bins are not guaranteed to contain sample points, in fact, bins are frequently empty in higher dimensional discretizations. One could extensively investigate strategies that describe how to handle these occurrences, however, this is beyond the scope of the current study. Here we simply let empty bins be described by the closest, in Euclidean sense, nonempty bin. In the occurrence of multiple closest bins, our implementation chooses the first closest bin listed in the storage format of the data set. While this is an *ad hoc* choice, we stress that with our chosen sample size M and bin size N_B (see Table 2.1), this is an extremely rare occurrence. This did not occur at all in most of our experiments; in the worst case ($C=4$, see Section 3.3) it affected only 0.01% of the reduced model time steps. However, this could be a point of improvement in future work.

In Figure 2.2, we show the simple algorithm used to integrate the reduced heat bath model (2.2) over time. In the following sections, we discuss the stochastic methods that describe the temporal evolution of $\tilde{r}$.

3. Numerical methods

3.1. Empirical distribution. In this section, we discuss the method of sampling $\tilde{r}$ directly from the sample data's *empirical* distribution, as formally defined in (3.1). This strategy has an obvious limitation, in that it can only sample from the values of r observed in the fully resolved simulation. However, for a stationary process, this empirical distribution of r conditioned on past values (see Section 1.2) will converge to the exact joint distribution in the limit of infinite data. Basic experiments show that simulations sampling instead from an unconditioned empirical distribution are highly unstable.

input: $\boldsymbol{Q}$: vector of sample data for q, length M.
$\boldsymbol{P}$: vector of sample data for p, length M.
$\boldsymbol{R}$: vector of sample data for r, length M.
c_i : set of conditioning variables, size C.
α_b : C-dimensional bins, for all $1 \leqslant b \leqslant B$.
$\min(\alpha_b)$: vector of minimum values per dimension over all α_b, length C.
$\text{step}(\alpha_b)$: vector of bin size per dimension, length C.
method : the stochastic approach used to approximate $\tilde{r}$, options: empirical, CMC, bin-wise OU and linear OU.

$(\tilde{q}_0, \tilde{p}_0, \tilde{r}_0) = (q_0, p_0, r_0)$
$i = 0$
for $i := 0$ to $N-1$ **do**
    /* Update $\tilde{q}$ and $\tilde{p}$ */
    $\tilde{p}_{i+1} = \tilde{p}_i - \Delta\tau\ V'(\tilde{q}_i) + \Delta\tau\ G^2(\tilde{r}_i - J\tilde{q}_i)$
    $\tilde{q}_{i+1} = \tilde{q}_i - \Delta\tau\ \tilde{p}_{i+1}$

    /* Find the bin number b such that $\tilde{c}_i \in \alpha_b$ */
    $b = \lceil \tilde{c}_i - \min(\alpha_b) \rceil ./ \text{step}(\alpha_b)$

    /* Update $\tilde{r}$ by random sampling */
    $\tilde{r}_{i+1} \sim \text{distr}(\text{method}, b)$
endfor

FIG. 2.2. *Algorithm for the time integration of the reduced model for a given set of conditioning variables c and stochastic approach.*

3.1.1. Reproducing statistical moments of distinguished particle. Let us define $\mathcal{U}(\rho_b)$ to denote the uniform distribution on the discrete set ρ_b, i.e. if $U \sim \mathcal{U}(\rho_b)$ then U has equal probability of being any element of the set ρ_b. The empirical approach fits the conditional residual term $\tilde{r}$ to r as follows:

$$\tilde{r}_{i+1} \sim \mathcal{U}(\rho_b), \quad \text{where } b : \tilde{c}_i \in \alpha_b. \tag{3.1}$$

Since q_i and r_{i+1} show a strong correlation, let us consider sampling $\tilde{r}_{i+1}$ from the distribution of $r_{i+1} | q_i$. We integrate the reduced model by using the algorithm in Figure 2.2 and (3.1) with $c_i = \{q_i\}$, and compare the resulting distributions of $\tilde{p}$ and $\tilde{q}$ to those of the fully resolved p and q. Each of the distributions is plotted in Figure 3.1.

Figure 3.1 shows that sampling from the distribution in (3.1) is effective in that the general shape of the distributions is reproduced, but there is also clearly room for improvement, e.g., one notices an underestimated standard deviation for both $\tilde{q}$ and $\tilde{p}$. As suggested in Section 1.2, one expects better results when expanding the set of conditioning variables c_i. Therefore, let us compare the previous approach to the conditioned distribution of $r_{i+1} | q_i, r_i$. To clearly illustrate the differences, we plot the absolute error of the resulting distributions in Figure 3.2.

Figure 3.2 shows that the distributions of $\tilde{p}$ and $\tilde{q}$ for $c_{i,1} := \{q_i\}$ are improved upon greatly by $c_{i,2} := \{q_i, r_i\}$. As suggested in Section 1.2, the first four sample moments of

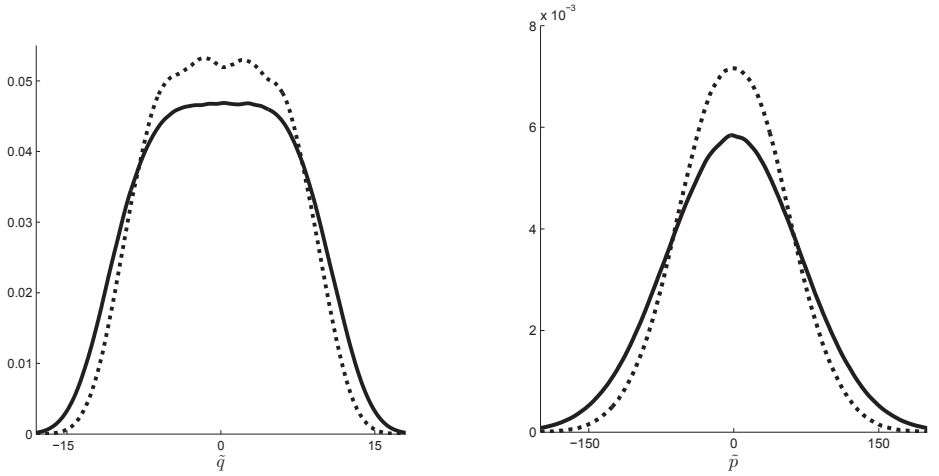

FIG. 3.1. *The distributions for positions $q,\tilde{q}$ (left) and momenta $p,\tilde{p}$ (right). The conditioned empirical distributions approximate sampling from $r_{i+1}|q_i$. A comparison between the distributions resulting from the reduced model (dotted lines) and resolved model (solid lines).*

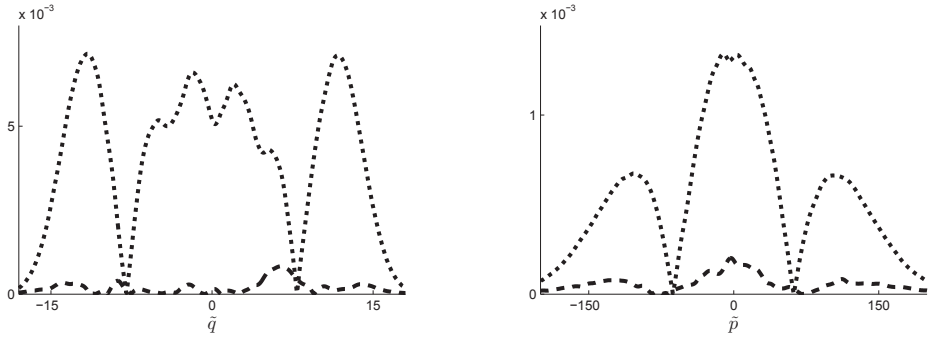

FIG. 3.2. *Absolute errors of the distributions for positions (left) and momenta (right). The conditioned empirical distributions approximate sampling from $r_{i+1}|c_i$. The absolute errors of both $c_{i,1}=\{q_i\}$ (dotted) and $c_{i,2}=\{q_i,r_i\}$ (dashed) are plotted.*

q and p, along with those of $\tilde{q}$ and $\tilde{p}$ for several cases are compared in Table 3.1. From this table one can conclude that conditioning on $c_{i,2}$ provides an overall improvement to $c_{i,1}$, the major improvement being the accuracy of the standard deviation for both $\tilde{q}$ and $\tilde{p}$, but also the kurtosis is more accurately reproduced. Since both q_i and r_i show a clear correlation with r_{i+1}, these results are expected. However, neither of the conditioning parameters improves the temporal correlation, as both condition on the same time step i. This is clearly shown in the autocorrelation functions plotted in Figure 3.3, where both of the approximations produce an inaccurate autocorrelation function. Because these procedures condition on specific time steps, the autocorrelation functions are dependent on the size of $\Delta\tau$, the integration time step of the reduced simulation; simulations discussed here use the parameter values as shown in Table 2.1.

3.1.2. Reproducing autocorrelation of distinguished particle. Our strategy for improving the autocorrelation function is to build more temporal correlation into the conditioning, i.e., we condition r_{i+1} on system variables from previous time steps.

$\boldsymbol{x}_i$	mean $\gamma_1(\boldsymbol{x}_i)$	std.dev. $\gamma_2(\boldsymbol{x}_i)$	skewness $\gamma_3(\boldsymbol{x}_i)$	kurtosis $\gamma_4(\boldsymbol{x}_i)$
p_i (reference)	0.00	68.4	$3.7\cdot10^{-4}$	3.00
$\tilde{p}_i$ $(c_{i,1}=\{q_i\})$	0.00	54.2	$-8.6\cdot10^{-4}$	2.96
$\tilde{p}_i$ $(c_{i,2}=\{q_i,r_i\})$	0.00	70.2	$-1.8\cdot10^{-3}$	3.00
$\tilde{p}_i$ $(c_{i,3}=\{q_i,r_i,r_{i-1}\})$	0.00	68.6	$1.5\cdot10^{-4}$	3.02
q_i (reference)	0.01	6.83	$-5.5\cdot10^{-3}$	2.18
$\tilde{q}_i$ $(c_{i,1}=\{q_i\})$	0.00	6.04	$-0.3\cdot10^{-3}$	2.16
$\tilde{q}_i$ $(c_{i,2}=\{q_i,r_i\})$	-0.01	6.86	$-0.5\cdot10^{-3}$	2.19
$\tilde{q}_i$ $(c_{i,3}=\{q_i,r_i,r_{i-1}\})$	0.02	6.78	$-4.8\cdot10^{-3}$	2.19

TABLE 3.1. *Sample moments for empirical approximations.*

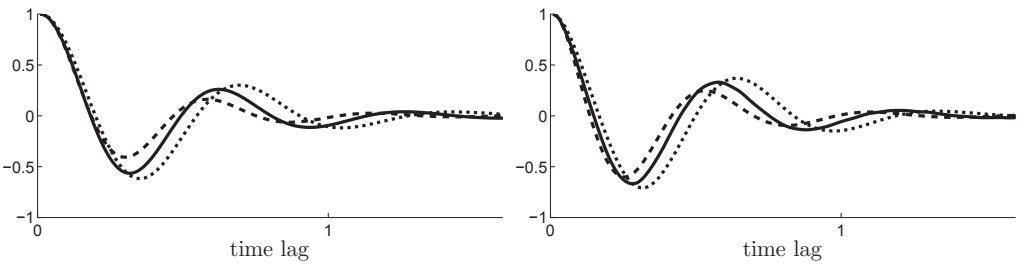

FIG. 3.3. *Autocorrelation functions for positions (left) and momenta (right). The conditioned empirical distributions approximate sampling from $r_{i+1}|c_i$. The autocorrelations for both $c_{i,1}=\{q_i\}$ (dotted lines) and $c_{i,2}=\{q_i,r_i\}$ (dashed lines) are plotted against the resolved autocorrelations (solid lines).*

As comparison to the results in Section 3.1.1, let us sample $\tilde{r}_{i+1}$ from the distribution of $r_{i+1}|c_{i,3}$, with $c_{i,3}=\{q_i,r_i,r_{i-1}\}$. Both the probability distributions of the approximated $\tilde{p}$ and $\tilde{q}$, as well as the associated autocorrelation functions are shown in Figure 3.4. As can be seen, they resemble the distributions and autocorrelations of the fully resolved model very closely. One can conclude that adding a greater dependence on the history of the sample data is greatly beneficial for approximating the autocorrelation function. Also, the sample moments of the reduced model variables remain comparable in quality (for $\tilde{q}$) or even improve (for $\tilde{p}$), see Table 3.1.

3.2. Conditional Markov chain approach. A natural evolution from the empirical approach, as described in Section 3.1, is to attempt to fit a continuous stochastic process to the sample data of r. The empirical approach will likely not perform to specification, because the empirical distribution samples exclusively from previously observed discrete values. This is especially true in situations where one cannot be convinced that the sample data is sufficiently representative of the entire range of possible values. In this section, we discuss how to use conditional Markov chains (CMCs) to model the stochastic process, similar (but not identical) to the approach from [3] and [4] (see also [12]).

3.2.1. Definition of the CMC. Expanding on the ideas put forward in [3], we define a CMC in which $\tilde{r}$ switches randomly between K deterministic functions f_k, with $1\leqslant k\leqslant K$. These functions describe the strong correlation between q and r and is such that $r_i=f_{k_i}(q_i)$, where $k_i=k(t_i)$ denotes the index of the specific function f in the ith

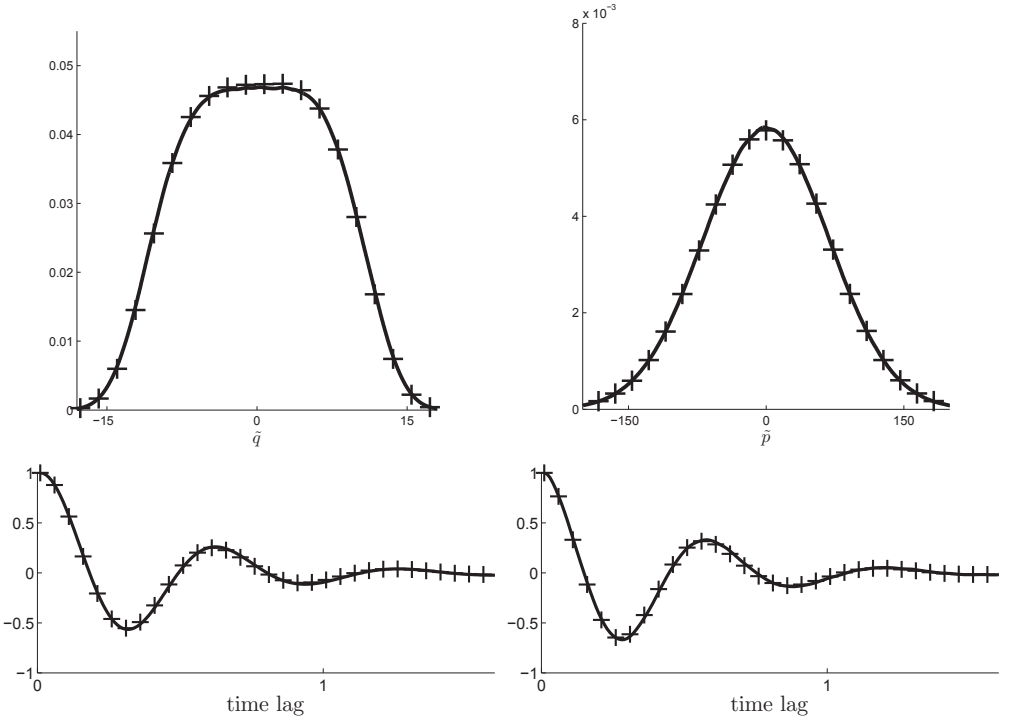

FIG. 3.4. *Distributions (top) and autocorrelation functions (bottom) for positions (left) and momenta (right). The conditioned empirical distributions are sampled from $r_{i+1}|q_i,r_i,r_{i-1}$. A comparison between the distributions and autocorrelations resulting from the reduced model (marked by +) and from the resolved model (solid lines).*

time step. Importantly, this method constructs $\tilde{r}$ as a piece-wise (in time) deterministic variable, therefore, one approximates transition distributions for $k_{i+1}|c_i$ rather than distributions of the form $r_{i+1}|c_i$. The numerical integration steps for a reduced model driven by a CMC residual term are defined as:

$$\begin{aligned} \tilde{p}_{i+1} = \tilde{p}_i - \ \Delta\tau V'(\tilde{q}_i) + \Delta\tau\ G^2(\tilde{r}_i - J\tilde{q}_i), \quad \tilde{q}_{i+1} = \tilde{q}_i + \Delta\tau\ \tilde{p}_{i+1}, \\ \tilde{k}_{i+1} \sim k_{i+1}|c_i = \tilde{c}_i, \quad \tilde{r}_{i+1} = f_{\tilde{k}_{i+1}}(\tilde{q}_{i+1}). \end{aligned} \tag{3.2}$$

We take linear functions f_{k_i}. An illustration of such functions fitted over a (q,r)-scatter plot is shown in Figure 3.5.

The conditioning variables c_i contain both model variables (e.g. q_i) and indices (e.g. k_i). The model variables are continuous, so they are binned as described in Section 2.3. Although many choices for c_i are possible, here we consider two sets $c_{i,3} = \{q_i, q_{i+1}, k_i\}$ and $c_{i,4} = \{q_i, q_{i+1}, k_i, k_{i-1}\}$. We emphasize that $c_{i,3}$ and $c_{i,4}$ are not implicit conditioning sets, because $\tilde{q}_{i+1}$ is calculated before $\tilde{r}_{i+1}$ is updated (see (3.2)). As k_i can take integer values ranging from 1 to K, the transition from k_i to k_{i+1} is governed by a set of $(K \times K)$ transition probability matrices in the case of $c_{i,3}$, one matrix for every bin α_b. There are $B = (N_B)^C$ bins in total, where C is the number of continuous variables in c_i ($C = 2$ for $c_{i,3}$ and $c_{i,4}$). With $c_{i,4}$, there are BK transition probability matrices of size $(K \times K)$, due to the additional conditioning on k_{i-1}.

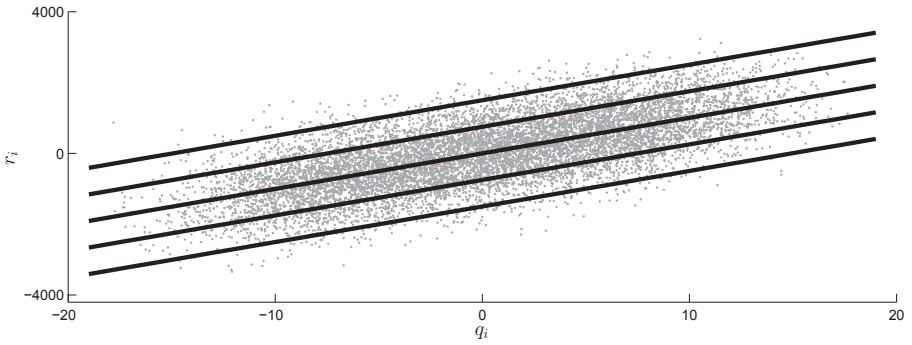

FIG. 3.5. *Example of five linear functions f_k fitted over the scatter plot of q_i vs. r_i.*

3.2.2. Numerical results. To approximate the bin-wise transition probabilities one first applies the mapping $(q_i,r_i)\to(q_i,k_i)$ to all data points, where $k_i := \arg\min_k |r_i - f_k(q_i)|$, i.e. k_i is chosen so that f_{k_i} is the function with minimal distance to the point (q_i,r_i) in the r-direction. After applying this mapping, one can easily count occurrences of transition paths in the sample data.

Constructing the transition probability matrices in this manner implies that k_{i+1} is dependent on all of k_i, q_i, and q_{i+1}. This has as effect that, for correct usage of these transition probabilities in the reduced model, the conditioning variables should at least include q_i, q_{i+1}, and k_i. In fact, we found that simulations where c_i does not include all three of these are often unstable.

Figure 3.6 compares the reduced model results of the simulations with conditioning variables $c_{i,3} = \{q_i, q_{i+1}, k_i\}$ and $c_{i,4} = \{q_i, q_{i+1}, k_i, k_{i-1}\}$. The conditioning variable k_{i-1} added in $c_{i,4}$ significantly improves the reproduced autocorrelation functions, similar to the results of the empirical distribution in Section 3.1.2.

The sample moments of the resolved simulation and the reduced simulations are shown in Table 3.2. This table shows that the conditioning parameters $c_{i,3}$ give a better approximation of moments of q and p than $c_{i,4}$, although with $c_{i,4}$ the autocorrelation functions are reproduced more accurately. Because, in Section 1.2, we posed that additional conditional variables to the distribution of $\tilde{r}$ should result in increased accuracy of the reduced model, this result is unexpected. However, a large number of parameters must be estimated to approximate the distribution of $k_{i+1}|c_i$. We recall the following definitions: C and D are the number of continuous and discrete variables in c_i, $B = (N_B)^C$ is the total number of bins, and K is the number of different functions $f_k(q)$. The number of parameters to be estimated for the CMC approach conditioning on a set of variables c_i is given by $(N_B)^C K^{D+1}$.

For the results in Figure 3.6 and Table 3.2 we used $K = 9$ and $B = 100$ (10×10 bins for q_i and q_{i+1} combined). This results in 8100 parameters when using $c_{i,3}$ and 72900 parameters when using $c_{i,4}$. This exponential scaling of the number of parameters is the bottleneck of the CMC approach: even for relatively simple problems it requires a very large data set to approximate all transition probabilities accurately.

Due to the described stability issues and exponential scaling of the number of parameters we choose not to pursue the CMC approaches any further here. Instead, in the next section we explore the use of a continuous-in-space stochastic process, so that the number of parameters remains minimal.

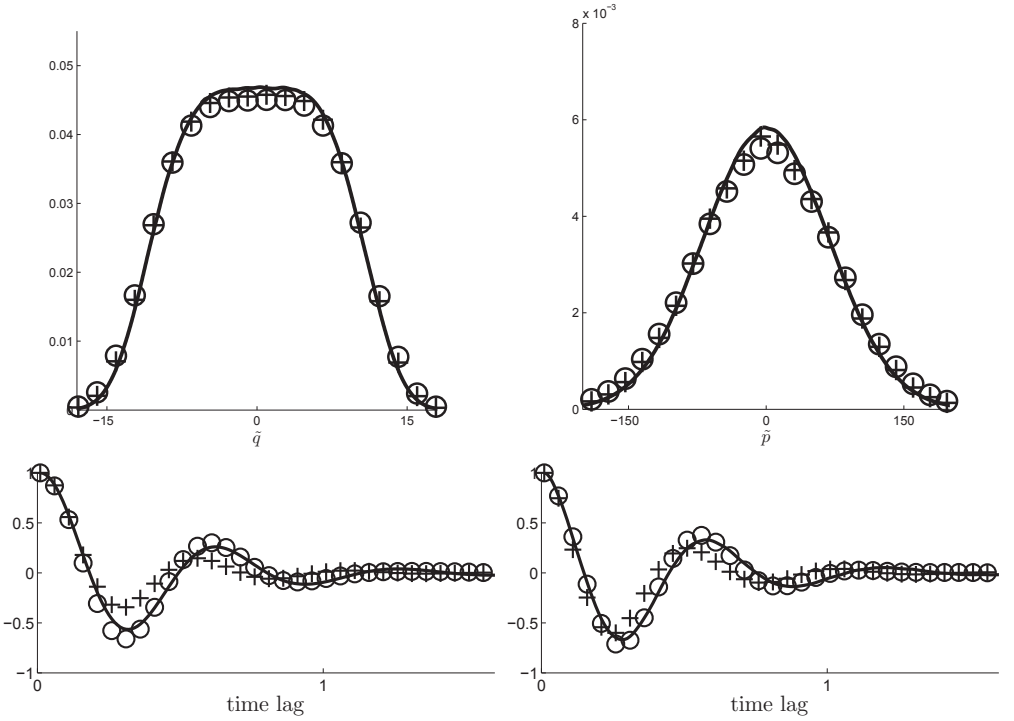

FIG. 3.6. *Distributions (top) and autocorrelation functions (bottom) for positions (left) and momenta (right). The CMC approach approximates sampling from $r_{i+1}|c_i$. A comparison between the distributions and autocorrelations resulting from the reduced models for $c_{i,3}=\{q_i,q_{i+1},k_i\}$ (marked by +) and for $c_{i,4}=\{q_i,q_{i+1},k_i,k_{i-1}\}$ (marked by ○), and from the resolved model (solid lines).*

$\boldsymbol{x}_i$	#params.	mean $\gamma_1(\boldsymbol{x}_i)$	std.dev. $\gamma_2(\boldsymbol{x}_i)$	skewness $\gamma_3(\boldsymbol{x}_i)$	kurtosis $\gamma_4(\boldsymbol{x}_i)$
p_i (reference)	–	0.00	68.4	$3.7\cdot10^{-4}$	3.00
$\tilde{p}_i$ ($c_{i,3}=\{q_i,q_{i+1},k_i\}$)	8100	0.00	71.8	$1.2\cdot10^{-3}$	3.00
$\tilde{p}_i$ ($c_{i,4}=\{q_i,q_{i+1},k_i,k_{i-1}\}$)	72900	0.00	74.3	$-3.4\cdot10^{-4}$	3.02
q_i (reference)	–	0.01	6.83	$-5.5\cdot10^{-3}$	2.18
$\tilde{q}_i$ ($c_{i,3}=\{q_i,q_{i+1},k_i\}$)	8100	0.00	7.00	$-3.4\cdot10^{-3}$	2.18
$\tilde{q}_i$ ($c_{i,4}=\{q_i,q_{i+1},k_i,k_{i-1}\}$)	72900	0.00	7.11	$-2.8\cdot10^{-3}$	2.19

TABLE 3.2. *Sample moments for the CMC approximations.*

3.3. Ornstein–Uhlenbeck process. As discussed in Section 3.2.2, the CMC strategy requires a very large number of estimated parameters. In this section we present a stochastic representation that reduces the number of parameters significantly. Let us assume that the evolution of r can be approximated by the following Ornstein–Uhlenbeck (OU) process:

$$\dot{r} = -\theta(r-\mu) + \sigma\dot{W},$$

with Wiener process W and unknown parameters μ, θ, and σ. The evolution of r, as observed from the full model, is then used to approximate an OU process $\tilde{r}$ defined by:

$$\dot{\tilde{r}} = -\hat{\theta}(\tilde{r} - \hat{\mu}) + \hat{\sigma}\dot{W}. \tag{3.3}$$

The parameters $\hat{\boldsymbol{\theta}} := (\hat{\mu}, \hat{\theta}, \hat{\sigma})$ in (3.3) approximate the OU parameters $\boldsymbol{\theta} := (\mu, \theta, \sigma)$, thus implicitly fitting $\tilde{r}$ to r. In the following sections we discuss different methods for defining these OU estimators. We start in Section 3.3.1 with constant $\hat{\boldsymbol{\theta}}$ (i.e., independent of c_i), whereas in later sections we let $\hat{\boldsymbol{\theta}}$ depend on c_i.

3.3.1. Unconditional parameters. Introduce the notations $R_c = \sum_{i=1}^{M} r_i$, $R_m = \sum_{i=1}^{M} r_{i-1}$, $R_{cc} = \sum_{i=1}^{M} r_i^2$, $R_{mm} = \sum_{i=1}^{M} r_{i-1}^2$, and $R_{cm} = \sum_{i=1}^{M} r_i r_{i-1}$. The subscripts c and m are chosen to denote *current* and *minus*, respectively. Then, assuming a zero-limit of the sampling interval δt, the *standard* discrete-in-time estimators $\hat{\boldsymbol{\theta}}^{\mathrm{st}} := (\hat{\mu}^{\mathrm{st}}, \hat{\theta}^{\mathrm{st}}, \hat{\sigma}^{\mathrm{st}})$ for the OU parameters are given by [13]:

$$\begin{aligned} \hat{\mu}^{\mathrm{st}} &= M^{-1} R_c, \\ \hat{\theta}^{\mathrm{st}} &= \frac{R_{mm} - R_{cm} - \hat{\mu}^{\mathrm{st}}(R_m - R_c)}{\delta t (R_{mm} - 2\hat{\mu}^{\mathrm{st}} R_m + M(\hat{\mu}^{\mathrm{st}})^2)}, \\ (\hat{\sigma}^{\mathrm{st}})^2 &= M^{-1} \delta t^{-1} (R_{cc} - 2R_{cm} + R_{mm}). \end{aligned} \tag{3.4}$$

Sometimes, however, a small δt cannot be guaranteed because of run-time requirements, or a small δt is undesired [13]. If δt is not small, the estimators in (3.4) are biased. Therefore, let us also consider the more exact maximum likelihood (ML) estimators $\hat{\boldsymbol{\theta}}^{\mathrm{ex}} := (\hat{\mu}^{\mathrm{ex}}, \hat{\theta}^{\mathrm{ex}}, \hat{\sigma}^{\mathrm{ex}})$, as discussed in, e.g, [17]. By omitting the assumption $\delta t \to 0$ and using the Markovian nature of the OU process, these exact ML estimators follow from maximizing the log-likelihood function:

$$\log \mathcal{L}(\hat{\boldsymbol{\theta}}^{\mathrm{ex}} \,|\, \boldsymbol{R}) = \log P(r_0 \,|\, \hat{\boldsymbol{\theta}}^{\mathrm{ex}}) + \sum_{i=1}^{M} \log P(r_i \,|\, r_{i-1}, \hat{\boldsymbol{\theta}}^{\mathrm{ex}}). \tag{3.5}$$

Making the additional assumption that the sample data is stationary, we know:

$$r_i \,|\, r_{i-1}, \hat{\boldsymbol{\theta}}^{\mathrm{ex}} \sim \mathcal{N}\left(r_{i-1}\eta + \hat{\mu}^{\mathrm{ex}}(1-\eta), (\zeta\hat{\sigma}^{\mathrm{ex}})^2\right),$$

where $\eta := \exp(-\hat{\theta}^{\mathrm{ex}}\delta t)$ and $\zeta^2 := (2\hat{\theta}^{\mathrm{ex}})^{-1}(1-\eta^2)$.

We assume the distribution of r_0 does not depend on $\hat{\boldsymbol{\theta}}$. Therefore, we ignore the term $P(r_0 \,|\, \hat{\boldsymbol{\theta}}^{\mathrm{ex}})$ for the maximization of (3.5). Substituting the conditional probabilities and removing the conditional distribution $P(r_0 \,|\, \hat{\boldsymbol{\theta}}^{\mathrm{ex}})$ from (3.5) results in the following log-likelihood:

$$\begin{aligned} &\log \mathcal{L}(\hat{\boldsymbol{\theta}}^{\mathrm{ex}} \,|\, \boldsymbol{R}) \approx \sum_{i=1}^{M} \log P(r_i \,|\, r_{i-1}, \hat{\boldsymbol{\theta}}^{\mathrm{ex}}) \\ &= -\frac{M}{2}\log(2\pi) - M\log(\zeta\hat{\sigma}^{\mathrm{ex}}) - \frac{1}{2(\zeta\hat{\sigma}^{\mathrm{ex}})^2} \sum_{i=1}^{M} \left(r_i - r_{i-1}\eta - \hat{\mu}^{\mathrm{ex}}(1-\eta)\right)^2. \end{aligned} \tag{3.6}$$

By maximizing (3.6) with respect to each of the parameters, the exact ML estima-

tors are found to equal:

$$\begin{aligned}
\hat{\mu}^{\mathrm{ex}} &= \frac{R_c R_{mm} - R_m R_{cm}}{M(R_{mm} - R_{cm}) - R_m^2 + R_c R_m}, \\
\hat{\theta}^{\mathrm{ex}} &= -\delta t^{-1} \log \frac{R_{cm} - \hat{\mu}^{\mathrm{ex}}(R_c + R_m) + M(\hat{\mu}^{\mathrm{ex}})^2}{R_{mm} - 2\hat{\mu}^{\mathrm{ex}} R_m + M(\hat{\mu}^{\mathrm{ex}})^2}, \\
(\hat{\sigma}^{\mathrm{ex}})^2 &= 2\hat{\theta}^{\mathrm{ex}} M^{-1}(1-\eta^2)^{-1} \ (\ R_{cc} - 2\eta R_{cm} + \eta^2 R_{mm} \\
&\qquad - 2\hat{\mu}^{\mathrm{ex}}(R_c - \eta R_m)(1-\eta) + M(\hat{\mu}^{\mathrm{ex}})^2(1-\eta)^2 \ ).
\end{aligned} \tag{3.7}$$

These estimators are equivalent to the standard ML estimators (3.4) if one assumes the limits $\delta t \to 0$ and $M \to \infty$ (see Appendix A). Note that the exact ML estimators (3.7) can be calculated sequentially from sample data.

Next, let us compare the quality of the respective methods by fitting both sets of estimators to sample data generated by a reference OU process with known parameters. Because both $\hat{\mu}^{\mathrm{st}}$ and $\hat{\mu}^{\mathrm{ex}}$ are independent of δt, we only compare approximations for σ and θ. Both the standard and exact ML estimators, fitted to this reference process, are shown in Figure 3.7. This figure shows that the standard ML estimators (3.4) indeed become strongly biased as δt increases, whereas the exact ML estimators (3.7) remain very accurate up to at least δt values of 1.5, where sampling error starts to be an issue. Therefore, the exact ML estimators are the clear choice for the rest of our experiments.

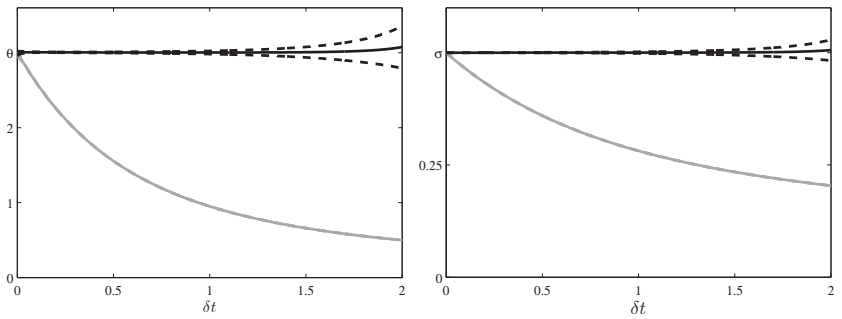

FIG. 3.7. *Mean (solid) and standard deviation (dashed) of the standard (gray) and exact (black) ML estimators, in (3.4) and (3.7) respectively, for a reference OU process with* $(\mu,\sigma,\theta) = (1, 0.5, 3)$. *The estimates plotted for each sampling interval* δt *are averages over 100 independent OU simulations with the given parameters. Each OU simulation stores* 10^6 *data points, where a data point is saved after* 100 *time steps of the reference process. The sampling interval of the OU simulations is* 10^{-3}. *We test the estimators as* δt *ranges from* 10^{-3} *to* 2, *in increments of* 10^{-3}. *This causes the growing sampling error shown as* $\delta t \to 2$. *Note that while the standard deviation of the standard ML estimators (gray dotted lines) is plotted in the figures, these dotted lines lie too close to the standard ML estimator mean to be visible.*

3.3.2. Conditional parameters with binning. We now generalize the methods from Section 3.3.1 to be in line with those in sections 3.1 and 3.2 by conditioning the OU parameters (and thus the process $\tilde{\boldsymbol{R}}$), on the model variables c. Building on the binning strategy, as explained in Section 2.3, we define estimators $\hat{\boldsymbol{\theta}}^{\mathrm{pc}} := (\hat{\mu}^{\mathrm{pc}}, \hat{\theta}^{\mathrm{pc}}, \hat{\sigma}^{\mathrm{pc}})$ that are piece-wise constant in c_i. It must be mentioned that this approach implicitly relies on small δt because the piece-wise constant assumption.

The c_i-dependency, being piece-wise constant, can be included in the likelihood function. First, we introduce the following notation:

$$\hat{\mu}^{\mathrm{pc}}(c_i) := \hat{\mu}_b^{\mathrm{pc}}, \quad \hat{\theta}^{\mathrm{pc}}(c_i) := \hat{\theta}_b^{\mathrm{pc}}, \quad \hat{\sigma}^{\mathrm{pc}}(c_i) := \hat{\sigma}_b^{\mathrm{pc}}, \quad \text{if } c_i \in \alpha_b.$$

The parameters $\hat{\boldsymbol{\theta}}_b^{\mathrm{pc}} := (\hat{\mu}_b^{\mathrm{pc}}, \hat{\theta}_b^{\mathrm{pc}}, \hat{\sigma}_b^{\mathrm{pc}})$ can be calculated by restricting the estimators (3.7) to the sample data points that lie in α_b. Note that we assume that r_i is only dependent on c_i, and not on $c_{i'}$ with $i' < i$. Similar to (3.6), the log-likelihood function can now be written as,

$$\log \mathcal{L}(\hat{\boldsymbol{\theta}}^{\mathrm{pc}} \,|\, \boldsymbol{R}, \boldsymbol{C}) \approx \sum_{i=1}^{M} \log P(r_i \,|\, r_{i-1}, \hat{\boldsymbol{\theta}}_b^{\mathrm{pc}}), \quad \text{where } c_{i-1} \in \alpha_b. \tag{3.8}$$

Maximizing (3.8) over the parameters ($3B$ in total) is straightforward and leads to the following estimators for each of the bins:

$$\begin{aligned}
\hat{\mu}_b^{\mathrm{pc}} &= \frac{R_{b,c}R_{b,mm} - R_{b,m}R_{b,cm}}{|\rho_b|(R_{b,mm} - R_{b,cm}) - R_{b,m}^2 + R_{b,c}R_{b,m}}, \\
\hat{\theta}_b^{\mathrm{pc}} &= -\delta t^{-1} \log \frac{R_{b,cm} - \hat{\mu}_b^{\mathrm{pc}}(R_{b,c} + R_{b,m}) + |\rho_b|(\hat{\mu}_b^{\mathrm{pc}})^2}{R_{b,mm} - 2\hat{\mu}_b^{\mathrm{pc}} R_{b,m} + |\rho_b|(\hat{\mu}_b^{\mathrm{pc}})^2}, \\
(\hat{\sigma}_b^{\mathrm{pc}})^2 &= 2\hat{\theta}_b^{\mathrm{pc}} |\rho_b|^{-1} (1-\eta_b^2)^{-1} \ \big(\ R_{b,cc} - 2\eta_b R_{b,cm} + \eta_b^2 R_{b,mm} \\
&\qquad - 2\hat{\mu}_b^{\mathrm{pc}}(R_{b,c} - \eta_b R_{b,m})(1-\eta_b) + |\rho_b|(\hat{\mu}_b^{\mathrm{pc}})^2 (1-\eta_b)^2 \ \big),
\end{aligned} \tag{3.9}$$

where $|\rho_b|$ is the number of sample points in the bin α_b. Analogous to before, the following notations are used to restrict terms to a specific bin b: $\eta_b = \exp(-\hat{\theta}_b^{\mathrm{pc}} \delta t)$, $R_{b,c} = \sum_{i=1}^{M} r_i \mathbf{1}(c_{i-1} \in \alpha_b)$, $R_{b,m} = \sum_{i=1}^{M} r_{i-1} \mathbf{1}(c_{i-1} \in \alpha_b)$, $R_{b,cc} = \sum_{i=1}^{M} r_i^2 \mathbf{1}(c_{i-1} \in \alpha_b)$, $R_{b,mm} = \sum_{i=1}^{M} r_{i-1}^2 \mathbf{1}(c_{i-1} \in \alpha_b)$ and $R_{b,cm} = \sum_{i=1}^{M} r_i r_{i-1} \mathbf{1}(c_{i-1} \in \alpha_b)$.

Let us illustrate this approach by calculating the bin-wise estimators for the one-dimensional conditioning $r_{i+1} \,|\, q_i$. The stationary distribution of an OU process with parameters $(\hat{\mu}_b^{\mathrm{pc}}, \hat{\theta}_b^{\mathrm{pc}}, \hat{\sigma}_b^{\mathrm{pc}})$ is $\mathcal{N}(\hat{\mu}_b^{\mathrm{pc}}, (\hat{\sigma}_b^{\mathrm{pc}})^2 / 2\hat{\theta}_b^{\mathrm{pc}})$; the resulting mean and standard deviation for each bin are plotted over a (q,r) scatter plot in Figure 3.8.

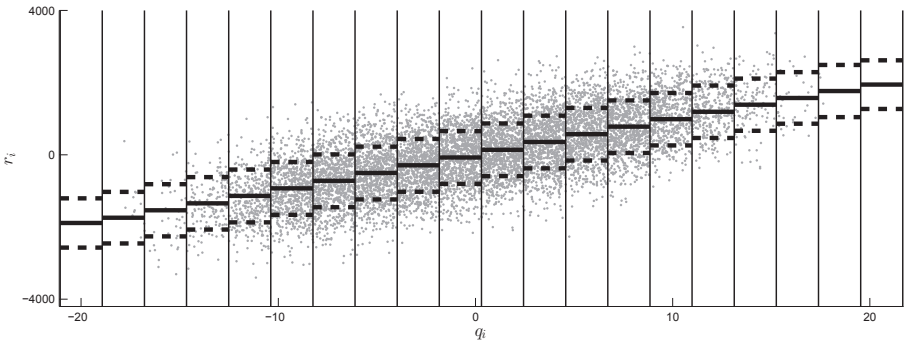

FIG. 3.8. *The mean (solid lines) and standard deviation (dotted lines) described by the stationary distribution of the OU estimators for each of the 20 bins approximating the distribution $r_{i+1} \,|\, q_i$. (Note that only 1% of the total number of data points used to obtain the estimators is shown in the plot.)*

3.3.3. Conditional parameters with a linearly fitted mean. In the specific case $\tilde{r}_{i+1} \sim r_{i+1} \,|\, q_i$, the means and standard deviations of the OU processes in the different bins are approximately linear (in q) and constant, respectively, as can be seen in Figure 3.8. In fact, our experiments show that the OU parameters themselves are either (approximately) constant ($\hat{\theta}_b^{\mathrm{pc}}$ and $\hat{\sigma}_b^{\mathrm{pc}}$), or linear in q ($\hat{\mu}_b^{\mathrm{pc}}$). This indicates that

we can reduce the total number of parameters significantly by using constant or linear functions of q. Thus, we define $(\hat{\mu}^{\mathrm{lf}}(q), \hat{\theta}^{\mathrm{lf}}(q), \hat{\sigma}^{\mathrm{lf}}(q))$ by

$$\hat{\mu}^{\mathrm{lf}}(q_i) = \hat{\mu}_0^{\mathrm{lf}} + \hat{\mu}_1^{\mathrm{lf}} q_i, \quad \hat{\theta}^{\mathrm{lf}}(q_i) = \hat{\theta}_0^{\mathrm{lf}}, \quad \hat{\sigma}^{\mathrm{lf}}(q_i) = \hat{\sigma}_0^{\mathrm{lf}}, \tag{3.10}$$

where $\hat{\boldsymbol{\theta}}^{\mathrm{lf}} := (\hat{\mu}_0^{\mathrm{lf}}, \hat{\mu}_1^{\mathrm{lf}}, \hat{\theta}_0^{\mathrm{lf}}, \hat{\sigma}_0^{\mathrm{lf}})$ is constant. When compared to the piece-wise constant OU estimators (3.9), this approach reduces the number of OU parameters from $3B$ to 4. Similar to (3.8), one can write the log-likelihood function for the parameters in (3.10) as,

$$\log \mathcal{L}(\hat{\boldsymbol{\theta}}^{\mathrm{lf}} \,|\, \boldsymbol{R}, \boldsymbol{Q}) \approx \sum_{i=1}^{M} P(r_i \,|\, r_{i-1}, q_{i-1}, \hat{\boldsymbol{\theta}}^{\mathrm{lf}}). \tag{3.11}$$

Analogous to Section 3.3.1, one obtains expressions for the estimators $\hat{\boldsymbol{\theta}}^{\mathrm{lf}}$ by maximizing (3.11) with respect to each of the parameters $\hat{\mu}_0^{\mathrm{lf}}$, $\hat{\mu}_1^{\mathrm{lf}}$, $\hat{\theta}_0^{\mathrm{lf}}$, and $\hat{\sigma}_0^{\mathrm{lf}}$ (see Appendix B for details on notation):

$$\begin{aligned}
\hat{\mu}_1^{\mathrm{lf}} &= -\frac{1}{3P_3}\left(P_2 + C + \frac{\Delta_0}{C}\right), \\
\hat{\mu}_0^{\mathrm{lf}} &= \frac{(\hat{\mu}_1^{\mathrm{lf}})^2 A_0 + \hat{\mu}_1^{\mathrm{lf}} B_0 + C_0}{\hat{\mu}_1^{\mathrm{lf}} D_0 + E_0} \\
\hat{\theta}_0^{\mathrm{lf}} &= -\delta t^{-1} \log\big((R_{mm} - 2\hat{\mu}_0^{\mathrm{lf}} R_m - 2\hat{\mu}_1^{\mathrm{lf}} X_{mm} + 2\hat{\mu}_0^{\mathrm{lf}}\hat{\mu}_1^{\mathrm{lf}} Q_m + M(\hat{\mu}_0^{\mathrm{lf}})^2 \\
&\qquad + (\hat{\mu}_1^{\mathrm{lf}})^2 Q_{mm})^{-1} (R_{cm} - \hat{\mu}_0^{\mathrm{lf}}(R_c + R_m) - \hat{\mu}_1^{\mathrm{lf}}(X_{cm} + X_{mm}) \\
&\qquad + 2\hat{\mu}_0^{\mathrm{lf}}\hat{\mu}_1^{\mathrm{lf}} Q_m + M(\hat{\mu}_0^{\mathrm{lf}})^2 + (\hat{\mu}_1^{\mathrm{lf}})^2 Q_{mm}) \big) \\
(\hat{\sigma}_0^{\mathrm{lf}})^2 &= 2\hat{\theta}_0^{\mathrm{lf}} M^{-1} (1 - \eta_0^2)^{-1} \big(R_{cc} - 2\eta_0 R_{cm} + \eta_0^2 R_{mm} \\
&\qquad + (1 - \eta_0)(2\hat{\mu}_0^{\mathrm{lf}}(\eta_0 R_m - R_c) + 2\hat{\mu}_1^{\mathrm{lf}}(\eta_0 X_{mm} - X_{cm})) \\
&\qquad + (1 - \eta_0)^2 (M(\hat{\mu}_0^{\mathrm{lf}})^2 + 2\hat{\mu}_0^{\mathrm{lf}}\hat{\mu}_1^{\mathrm{lf}} Q_m + (\hat{\mu}_1^{\mathrm{lf}})^2 Q_{mm}\ ).
\end{aligned} \tag{3.12}$$

The stationary distribution of the OU process with q fixed is given by $\mathcal{N}(\hat{\mu}_0^{\mathrm{lf}} + q\hat{\mu}_1^{\mathrm{lf}}, (\hat{\sigma}_0^{\mathrm{lf}})^2 / 2\hat{\theta}_0^{\mathrm{lf}})$; the resulting mean and standard deviation are plotted over a (q,r) scatter plot in 3.9.

3.3.4. Numerical results. As discussed in Section 2.3, it is possible that not all bins contain samples if they are equally sized. For the empirical approach in Section 3.1, this posed no serious problem. However, the accuracy of the estimated OU parameters is strongly affected if the sample size is too small. To keep the tests between methods comparable, we opt not to change the binning procedure, but instead opt to consider bins with less than 100 samples as empty.

Simulations that sample $\tilde{r}$ from the unconditioned distribution of r (using the constant ML estimators in (3.7)) are unstable. However, modeling $\tilde{r}$ as an OU process that is either piece-wise constant or linear in c_i (using the bin-wise or linearly fitted ML estimators, (3.9) or (3.12)) compares favorably to the previously discussed strategies. Whereas both the empirical and CMC approaches need 2 and 3 conditioning variables, respectively, to accurately reproduce the distributions of q and p, the reduced simulations using the conditioned OU process need only $c_{i,1} = \{q_i\}$ to reproduce these distributions very accurately. These results are illustrated in Figure 3.10 and presented with

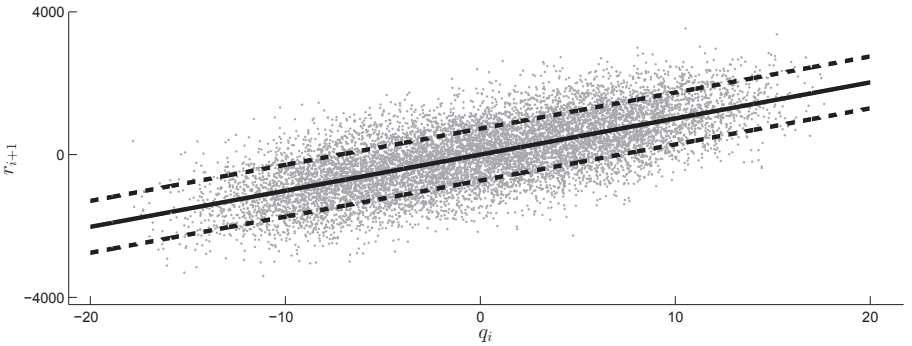

FIG. 3.9. *The mean and standard deviation of the stationary distribution described by linear OU parameters that approximate sampling $\tilde{r}_{i+1}$ from the distribution of $r_{i+1}|q_i$. (Note that only 1% of the total number of data points used to obtain the estimators is shown in the plot.)*

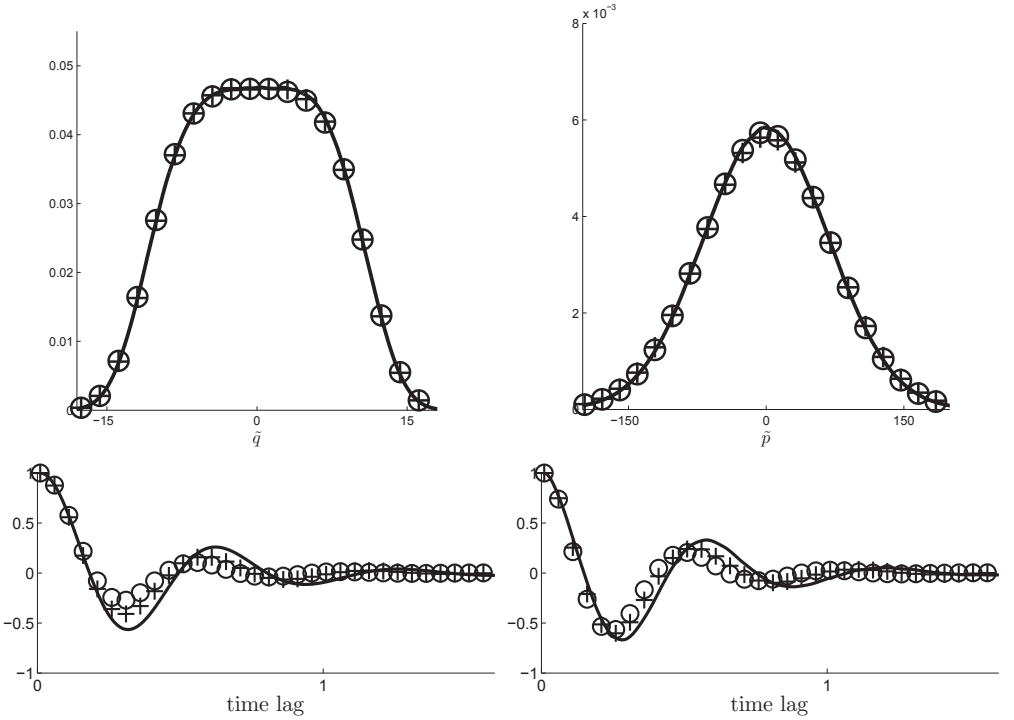

FIG. 3.10. *Distributions (top) and autocorrelation functions (bottom) for positions (left) and momenta (right). The applied OU approaches approximate sampling from $r_{i+1}|q_i$. A comparison between the distributions and autocorrelations resulting from the bin-wise (marked by +) and linearly fitted (marked by ○) ML estimators, and from the resolved model (solid lines).*

more detail in Table 3.3. The accurate reconstruction of the model variables' distributions is especially impressive for the OU parameters with linearly fitted mean (referred to from now on as *linearly fitted* OU parameters), because the linearly fitted OU process only uses 4 parameters, whereas the CMC approach and bin-wise OU approach need $(N_B)^C K^{D+1}$ (see Section 3.2.2) and $3(N_B)^C$ parameters respectively. However,

$\boldsymbol{x}_i$	#params.	mean $\gamma_1(\boldsymbol{x}_i)$	std.dev. $\gamma_2(\boldsymbol{x}_i)$	skewness $\gamma_3(\boldsymbol{x}_i)$	kurtosis $\gamma_4(\boldsymbol{x}_i)$
p_i (reference)	–	0.00	68.4	$3.7\cdot10^{-4}$	3.00
$\tilde{p}_i$ $(c_{i,1}=\{q_i\})$ linearly fitted	4	0.00	69.2	$4.5\cdot10^{-4}$	3.01
$\tilde{p}_i$ $(c_{i,1}=\{q_i\})$ bin-wise	30	0.00	70.3	$-1.2\cdot10^{-3}$	3.02
$\tilde{p}_i$ $(c_{i,2}=\{q_i,r_i\})$	300	0.00	72.1	$-2.6\cdot10^{-3}$	3.02
$\tilde{p}_i$ $(c_{i,3}=\{q_i,r_i,r_{i-1}\})$	3000	0.00	69.3	$-3.5\cdot10^{-3}$	2.99
q_i (reference)	–	0.01	6.83	$-5.5\cdot10^{-3}$	2.18
$\tilde{q}_i$ $(c_{i,1}=\{q_i\})$ linearly fitted	4	-0.01	6.87	$2.8\cdot10^{-3}$	2.18
$\tilde{q}_i$ $(c_{i,1}=\{q_i\})$ bin-wise	30	0.00	6.86	$3.6\cdot10^{-4}$	2.19
$\tilde{q}_i$ $(c_{i,2}=\{q_i,r_i\})$	300	-0.01	6.94	$-1.3\cdot10^{-4}$	2.19
$\tilde{q}_i$ $(c_{i,3}=\{q_i,r_i,r_{i-1}\})$	3000	0.01	6.82	$-1.8\cdot10^{-3}$	2.18

TABLE 3.3. *Sample moments for the OU approximations.*

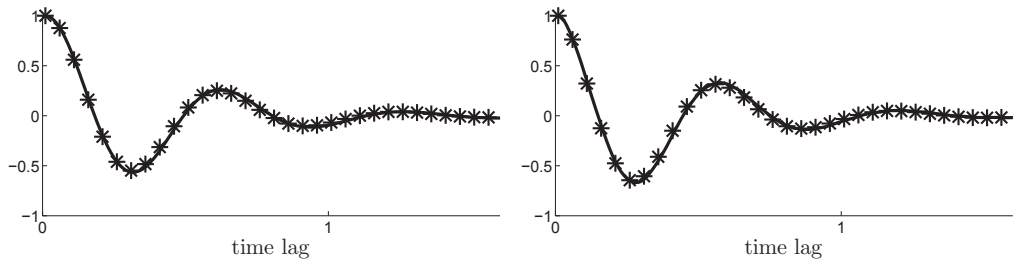

FIG. 3.11. *Autocorrelation functions for positions (left) and momenta (right). The piece-wise constant OU parameters are fitted to approximate sampling from* $r_{i+1}|q_i,r_i,r_{i-1}$. *A comparison between the autocorrelations resulting from the reduced model (marked by* $*$*) and from the resolved model (solid lines).*

as is the case with all other strategies, the autocorrelation function is reconstructed less accurately for $c_{i,1}$.

A downside of the linearly fitted OU parameters is that they are defined specifically for the case $c_{i,1}=\{q_i\}$. Generalization to other cases is nontrivial. The piece-wise constant OU parameters, however, can be easily conditioned on multiple variables. Similarly to the empirical approach, the resolved autocorrelation functions are approximated with high accuracy when the conditioning variables are extended to $c_{i,3}=\{q_i,r_i,r_{i-1}\}$, as shown in Figure 3.11.

4. Discussion

In this study we investigate how to use sample data, generated by a fully resolved multiscale model, to construct stochastic representations of unresolved processes in reduced models. We discuss three methods to model these stochastic representations, and tested the methods using the Kac–Zwanzig heat bath model. This heat bath model describes the dynamics of a distinguished particle, which is coupled linearly to a number of heat bath particles and moves over a potential. The stochastic methods aim

to model the dynamical effects of the heat bath particles to drive a reduced model that only resolves the distinguished particle. We compared the fully resolved model and the reduced models by the probability distributions, first four statistical moments and autocorrelation functions of the position q and momentum p of the distinguished particle.

In the reduced models, the sum of the positions of the heat bath particles, denoted r, is modeled as a stochastic process. This is done in three different ways: (i) sampling from the empirical (conditional) distribution of r, (ii) using a discrete Markov chain to switch between several functions $r = f_k(q)$, and (iii) modeling r as an OU process with q-dependent parameters. As mentioned before, the stochastic processes driving the reduced model were conditioned on the position of the distinguished particle. In some tests, the past state of r was also added to the set of conditioning variables. Extending the set of conditioning variables improves results, as is demonstrated most visibly in Section 3.1.2. We note that extending this set typically increases the number of parameters in the stochastic model, so that more data may be needed to estimate these accurately, see Section 3.2. Notwithstanding, with appropriate conditioning of the stochastic process for r, the distributions and autocorrelations of $\tilde{q}$ and $\tilde{p}$ in the reduced model resemble those of q and p of the fully resolved model very closely, see in particular figures 3.4 and 3.11.

The advantage of the empirical distribution approach over the other methods is that it is more robust if the available data set is rather small. The empirical distribution samples uniformly from the data, so that any nonempty sample set is, by construction, somewhat representative of the dynamics of r. However, this also restricts the empirical distribution sampling to the range of the data set, which might not be representative of the exact joint distribution of r for small data sets. In this approach, no parameters are estimated, the data only needs to be partitioned into bins. By contrast, the CMC and binned OU approaches are more sensitive to small data sets, because limited data affects the parameter estimates. These approaches involve a large number of parameters that must be estimated, most notably the CMC approach, see tables 3.2 and 3.3.

The linearly fitted OU approach reduces the number of parameters to 4, and is still able to reproduce the distributions of the resolved model variables very accurately. However, we note that extending this approach to one where the OU parameters θ and σ also have functional dependence on q (or some other conditioning variable) will be difficult, as will generalizations to nonlinear functional dependence.

As mentioned, the data needed for fitting the stochastic models for r come from a simulation of the fully resolved model. It may seem superfluous to formulate a reduced model if simulations with the full model are computationally feasible. However, if one wishes to simulate a multiscale system over a very long time interval, but fully resolved simulations are only feasible over a much shorter time interval, an efficient yet accurate reduced model can be very useful. Furthermore, in some cases it is possible to use data from observations instead of simulation data (see [5] for an example). In those situations, data-driven modeling approaches are also useful. Finally, for spatially extended systems such as atmospheric or oceanic flows, a fully resolved simulation may be only computationally feasible on part of the spatial domain of interest. The methods discussed in this study allow one to construct a spatially localized stochastic model for unresolved processes. By using identical yet independent copies of this local stochastic model, one can cover the entire spatial domain.

We emphasize that the computational gain of simulating with the reduced model instead of the fully resolved model can be very large. The Kac–Zwanzig heat bath

model as used in this study has 202 degrees of freedom (positions and momenta of the distinguished particle and all 100 heat bath particles), and is integrated with time step 10^{-4}. By contrast, the reduced model has 3 degrees of freedom $(\tilde{q},\tilde{p},\tilde{r})$ and integration time step 10^{-2}. An application example is [4], where the fully resolved model is a large eddy simulation (LES) model for atmospheric convection. The LES model in that study used $512\times512\times80$ gridpoints for spatial discretization, and an integration time step on the order of seconds. The CMC used to represent convection as simulated by the LES model contained 10 discrete states, with random switching between the discrete states at time steps of 1 minute.

In future work we aim to use the methods presented here for ocean circulation models. For example, the strategies described in [15] propose a covariate that correlates strongly to the residual term in reduced vorticity equations. Investigating how our methods can be applied to such models is an exciting topic for future study.

Appendix A. The equivalence of the exact and standard ML estimators in appropriate limits. The unconditional maximum likelihood (ML) estimators, as described in Section 3.3.1, are obtained by maximizing the log-likelihood in (3.6). Here we make a distinction between two types of ML estimators: the standard estimators $\hat{\boldsymbol{\theta}}^{\mathrm{st}}=(\hat{\mu}^{\mathrm{st}},\hat{\theta}^{\mathrm{st}},\hat{\sigma}^{\mathrm{st}})$ in (3.4), and the exact estimators $\hat{\boldsymbol{\theta}}^{\mathrm{ex}}=(\hat{\mu}^{\mathrm{ex}},\hat{\theta}^{\mathrm{ex}},\hat{\sigma}^{\mathrm{ex}})$ in (3.7). The standard ML estimators are obtained by imposing the limit $\delta t\downarrow 0$ on the log-likelihood equation, but the exact ML estimators make no such assumption. However, we show here that the exact estimators tend to the standard estimators as δt and the sample size M go to 0 and ∞, respectively.

Let us make the following assumptions about the model's sample data $\boldsymbol{R}=(r_0,r_1,\ldots,r_M)$:

1. $|r_i|<\infty$, for $0\leqslant i\leqslant M$.
2. $\bar{r}:=\mathbf{E}(r_i)\ \forall i$ (stationarity), and $|\bar{r}|<\infty$.
3. $\mathrm{Var}(r)=\mathbf{E}(r_i-\bar{r})^2\ \forall i$ (stationarity), and $0<\mathrm{Var}(r)<\infty$.
4. $\mathrm{ACF}:=(\mathbf{E}(r_ir_{i-1})-\bar{r}^2)\mathrm{Var}(r)^{-1}$, and $|\mathrm{ACF}|<1$. Note that ACF is dependent on δt, therefore let us also assume:
   (a) $\lim_{\delta t\downarrow 0}\mathrm{ACF}=1$.
   (b) $\lim_{\delta t\downarrow 0}(\delta t^{-1}(\mathrm{ACF}-1))=:-a$, with $0<a<\infty$. This essentially restricts the right derivative of the autocorrelation function from nearing infinite or zero as δt goes to 0.

To show that the standard and exact ML estimators are equivalent in the limits $M\to\infty$ and $\delta t\downarrow 0$, we will first consider each estimator in the limit $M\to\infty$. Let us therefore list the following known properties:

$$\begin{aligned}
&\lim_{M\to\infty}(M^{-1}R_c)=\bar{r},\\
&\lim_{M\to\infty}(M^{-1}R_m)=\bar{r},\\
&\lim_{M\to\infty}(M^{-1}R_{cc})=\bar{r}^2+\mathrm{Var}(r),\\
&\lim_{M\to\infty}(M^{-1}R_{mm})=\bar{r}^2+\mathrm{Var}(r),\\
&\lim_{M\to\infty}(M^{-1}R_{cm})=\mathbf{E}(r_ir_{i-1})=\bar{r}^2+\mathrm{Var}(r)\,\mathrm{ACF}.
\end{aligned}\tag{A.1}$$

Equivalence for $\hat{\mu}$. It follows directly from (A.1) that

$$\lim_{M\to\infty} \hat{\mu}^{\mathrm{st}} = \bar{r}. \tag{A.2}$$

Now, let us consider the estimator $\hat{\mu}^{\mathrm{ex}}$ (3.7):

$$\begin{aligned}
\hat{\mu}^{\mathrm{ex}} &= \frac{R_c R_{mm} - R_m R_{cm}}{M(R_{mm} - R_{cm}) - R_m^2 + R_c R_m} \\
&= \frac{R_m(R_{mm} - R_{cm}) + R_{mm}(r_M - r_0)}{M(R_{mm} - R_{cm}) + R_m(r_M - r_0)} \\
&= \frac{\frac{R_m}{M}\left(\frac{R_{mm}}{M} - \frac{R_{cm}}{M}\right) + \frac{R_{mm}}{M}\left(\frac{r_M - r_0}{M}\right)}{\left(\frac{R_{mm}}{M} - \frac{R_{cm}}{M}\right) + \frac{R_m}{M}\left(\frac{r_M - r_0}{M}\right)}.
\end{aligned}$$

Because we analyze $\hat{\mu}^{\mathrm{ex}}$ in the limit of $M \to \infty$, we first show that the denominator above does not go to 0 (using the properties in (A.1)):

$$\begin{aligned}
&\lim_{M\to\infty} \left(\left(\frac{R_{mm}}{M} - \frac{R_{cm}}{M}\right) + \frac{R_m}{M}\left(\frac{r_M - r_0}{M}\right) \right) \\
=&\bar{r}^2 + \mathrm{Var}(r) - \bar{r}^2 - \mathrm{Var}(r)\,\mathrm{ACF} \\
=&\mathrm{Var}(r)(1 - \mathrm{ACF}) > 0,
\end{aligned} \tag{A.3}$$

which allows us to split the limit. Thus

$$\begin{aligned}
\lim_{M\to\infty} \hat{\mu}^{\mathrm{ex}} &= \frac{\lim_{M\to\infty}\left(\frac{R_m}{M}\right)\lim_{M\to\infty}\left(\frac{R_{mm}}{M} - \frac{R_{cm}}{M}\right) + \lim_{M\to\infty}\left(\frac{R_{mm}}{M}\right)\lim_{M\to\infty}\left(\frac{r_M - r_0}{M}\right)}{\lim_{M\to\infty}\left(\frac{R_{mm}}{M} - \frac{R_{cm}}{M}\right) + \lim_{M\to\infty}\left(\frac{R_m}{M}\right)\lim_{M\to\infty}\left(\frac{r_M - r_0}{M}\right)} \\
&= \frac{\lim_{M\to\infty}\left(\frac{R_m}{M}\right)\lim_{M\to\infty}\left(\frac{R_{mm}}{M} - \frac{R_{cm}}{M}\right)}{\lim_{M\to\infty}\left(\frac{R_{mm}}{M} - \frac{R_{cm}}{M}\right)} \\
&= \lim_{M\to\infty}\left(\frac{R_m}{M}\right) \\
&= \bar{r}.
\end{aligned}$$

This together with (A.2) proves:

$$\lim_{M\to\infty} \hat{\mu}^{\mathrm{st}} = \lim_{M\to\infty} \hat{\mu}^{\mathrm{ex}}.$$

Equivalence for θ. Directly from (A.1) we see that

$$\lim_{M\to\infty} \delta t \hat{\theta}^{\mathrm{st}} = 1 - \mathrm{ACF},$$

and because δt and M are independent, we even see that:

$$\lim_{\delta t \downarrow 0}\left(\lim_{M\to\infty} \hat{\theta}^{\mathrm{st}}\right) = \lim_{\delta t \downarrow 0}\left(\delta t^{-1}(1 - \mathrm{ACF})\right) = a. \tag{A.4}$$

Next, let us consider the following operations on the estimator $\hat{\theta}^{\text{ex}}$ (3.7):

$$1-e^{-\hat{\theta}^{\text{ex}}\delta t} = \frac{R_{mm}-R_{cm}-\hat{\mu}^{\text{ex}}(R_m-R_c)}{R_{mm}-2\hat{\mu}^{\text{ex}}R_m+M(\hat{\mu}^{\text{ex}})^2} = \frac{\frac{R_{mm}}{M}-\frac{R_{cm}}{M}-\hat{\mu}^{\text{ex}}(\frac{R_m}{M}-\frac{R_c}{M})}{\frac{R_{mm}}{M}-2\hat{\mu}^{\text{ex}}\frac{R_m}{M}+(\hat{\mu}^{\text{ex}})^2}.$$

Let us first show that the above denominator does not tend to 0 in the limit of $M\to\infty$.

$$\lim_{M\to\infty}\left(\frac{R_{mm}}{M}-2\hat{\mu}^{\text{ex}}\frac{R_m}{M}+(\hat{\mu}^{\text{ex}})^2\right)=\text{Var}(r)>0. \tag{A.5}$$

This allows us to split the limit by using (A.3) and (A.5). Thus

$$\lim_{M\to\infty}\left(1-e^{-\hat{\theta}^{\text{ex}}\delta t}\right)=\frac{\text{Var}(r)(1-\text{ACF})}{\text{Var}(r)} =1-\text{ACF}, \tag{A.6}$$

which, since the logarithm function is continuous in $1-\text{ACF}$ (because $-1<\text{ACF}<1$), results in:

$$\lim_{M\to\infty}\hat{\theta}^{\text{ex}}=-\delta t^{-1}\log(\text{ACF}). \tag{A.7}$$

To evaluate (A.7) in the limit $\delta t\downarrow 0$, note that both $\log(\text{ACF})$ and δt go to 0 in the limit $\delta t\downarrow 0$, let us apply L'Hôspital's rule for one-sided limits:

$$\lim_{\delta t\downarrow 0}\left(\lim_{M\to\infty}\hat{\theta}^{\text{ex}}\right)=\lim_{\delta t\downarrow 0}\left(-\frac{d(\text{ACF})}{\text{ACF}d(\delta t)}\right) =\lim_{\delta t\downarrow 0}\left(-\delta t^{-1}(\text{ACF}-1)\right)=a. \tag{A.8}$$

Therefore, from (A.4) and (A.8), one concludes that:

$$\lim_{\delta t\downarrow 0}\lim_{M\to\infty}\hat{\theta}^{\text{st}}=\lim_{\delta t\downarrow 0}\lim_{M\to\infty}\hat{\theta}^{\text{ex}}=a. \tag{A.9}$$

Equivalence for σ. Directly from (A.1) we see that:

$$\lim_{M\to\infty}(\hat{\sigma}^{\text{st}})^2=2\delta t^{-1}(1-\text{ACF})\text{Var}(r).$$

Let us then use assumption 4(b) to arrive at:

$$\lim_{\delta t\downarrow 0}\lim_{M\to\infty}(\hat{\sigma}^{\text{st}})^2=2a\,\text{Var}(r). \tag{A.10}$$

Now, we recall from Section 3.3.1 the definition $\eta:=\exp(-\hat{\theta}^{\text{ex}}\delta t)$, and rewrite the estimator $\hat{\sigma}^{\text{ex}}$ (3.7) to:

$$\lim_{M\to\infty}(\hat{\sigma}^{\text{ex}})^2=\lim_{M\to\infty}(\frac{2\hat{\theta}^{\text{ex}}}{1-\eta^2}(\frac{R_{cc}}{M}-2\eta\frac{R_{cm}}{M}+\eta^2\frac{R_{mm}}{M} -2\hat{\mu}^{\text{ex}}(\frac{R_c}{M}-\eta\frac{R_m}{M})(1-\eta)+(\hat{\mu}^{\text{ex}})^2(1-\eta)^2)). \tag{A.11}$$

From (A.6) we know that

$$\lim_{M\to\infty} 1-\eta^2 = 1-\mathrm{ACF}^2 > 0,$$

which shows that the denominator in (A.11) does not go to 0 in the limit $M\to\infty$, therefore we can split up the limit and look at each of the elements separately. Rigorous algebraic operations lead to the simplified equation:

$$\lim_{M\to\infty} (\hat{\sigma}^{\mathrm{ex}})^2 = -2\delta t^{-1}\log(\mathrm{ACF})\mathrm{Var}(r),$$

to which we apply the limit $\delta t \downarrow 0$ (like in (A.8)) to finally obtain:

$$\lim_{\delta t\downarrow 0}\left(\lim_{M\to\infty} (\hat{\sigma}^{\mathrm{ex}})^2\right) = -2a\,\mathrm{Var}(r). \tag{A.12}$$

Therefore, from (A.10) and (A.12), one concludes:

$$\lim_{\delta t\downarrow 0}\lim_{M\to\infty} (\hat{\sigma}^{\mathrm{st}})^2 = \lim_{\delta t\downarrow 0}\lim_{M\to\infty} (\hat{\sigma}^{\mathrm{ex}})^2. \tag{A.13}$$

Appendix B. Explicit formulas for linearly fitted OU parameters. First, let us introduce the following notation: $\eta_0 := \exp(-\hat{\theta}_0^{\mathrm{lf}}\delta t)$, $Q_c = \sum_{i=1}^M q_i$, $Q_m = \sum_{i=1}^M q_{i-1}$, $Q_{cc} = \sum_{i=1}^M q_i^2$, $Q_{mm} = \sum_{i=1}^M q_{i-1}^2$, $Q_{cm} = \sum_{i=1}^M q_i q_{i-1}$, $X_{cc} = \sum_{i=1}^M r_i q_i$, $X_{mm} = \sum_{i=1}^M r_{i-1} q_{i-1}$, $X_{cm} = \sum_{i=1}^M r_i q_{i-1}$, $X_{mc} = \sum_{i=1}^M r_{i-1} q_i$ and

$$\begin{aligned}
A_0 &= Q_{mm}(R_c - R_m) - Q_m(X_{cm} - X_{mm}),\\
A_1 &= M(X_{cm} - X_{mm}) - Q_m(R_c - R_m),\\
B_0 &= R_m(X_{cm} + X_{mm}) + Q_m(R_{cm} - R_{mm}) - 2R_c X_{mm},\\
B_1 &= X_{mm}(R_c + R_m) - Q_m(R_{mm} - R_{cm}) - 2R_m X_{cm},\\
C_0 &= R_c R_{mm} - R_m R_{cm},\\
C_1 &= R_{mm} X_{cm} - R_{cm} X_{mm},\\
D_0 &= M(X_{cm} - X_{mm}) - Q_m(R_c - R_m)),\\
D_1 &= Q_{mm}(R_c - R_m) - Q_m(X_{cm} - X_{mm}),\\
E_0 &= M(R_{mm} - R_{cm}) + R_c R_m - R_m^2,\\
E_1 &= Q_{mm}(R_{mm} - R_{cm}) - X_{mm}(X_{mm} - X_{cm}).
\end{aligned}$$

An explicit formula for $\hat{\mu}_1^{\mathrm{lf}}$ is then given by the real root of the third order polynomial $(\hat{\mu}_1^{\mathrm{lf}})^3 P_3 + (\hat{\mu}_1^{\mathrm{lf}})^2 P_2 + \hat{\mu}_1^{\mathrm{lf}} P_1 + P_0$, where the polynomial coefficients are given by the following:

$$\begin{aligned}
P_3 &= A_0 A_1 B_0 + A_0 B_1 D_0 - A_0 D_1 E_0 - D_0^2 E_1\\
P_2 &= A_1 B_0^2 + 2A_0 A_1 C_0 + A_0 B_1 E_0 + B_0 B_1 D_0 + C_1 D_0^2 - B_0 D_1 E_0 - C_0 D_0 D_1 - 2D_0 E_0 E_1\\
P_1 &= 2A_1 B_0 C_0 + B_0 B_1 E_0 + B_1 C_0 D_0 + 2C_1 D_0 E_0 - C_0 D_1 E_0 - E_0^2 E_1\\
P_0 &= A_1 C_0 + B_1 C_0 E_0 + C_1 E_0^2.
\end{aligned}$$

After introducing the last notation $\Delta_0 = P_2^2 - 3P_3 P_1$, $\Delta_1 = 2P_2^2 - 9P_3 P_2 P_1 + 27P_3^2 P_0$ and $C = \sqrt[3]{\frac{\Delta_1 + \sqrt{\Delta_1^2 - 4\Delta_0^3}}{2}}$, one obtains the explicit formula for $\hat{\mu}_1^{\mathrm{lf}}$ and sequentially obtainable solutions for $\hat{\mu}_0^{\mathrm{lf}}$, $\hat{\theta}_0^{\mathrm{lf}}$ and $\hat{\sigma}_0^{\mathrm{lf}}$ in (3.12).

Acknowledgements. We thank Keith Myerscough for insightful discussions. This research is funded by the Netherlands Organisation for Scientific Research (NWO) through the Vidi project "Stochastic models for unresolved scales in geophysical flows".

REFERENCES

[1] P.S. Berloff, *On dynamically consistent eddy fluxes*, Dynam. Atmos. Oceans, 38(3), 123–146, 2005.
[2] P.S. Berloff, *Random-forcing model of the mesoscale oceanic eddies*, J. Fluid. Mech., 529, 71–95, 2005.
[3] D. Crommelin and E. Vanden-Eijnden, *Subgrid-scale parameterization with conditional Markov chains*, J. Atmos. Sci., 65(8), 2661–2675, 2008.
[4] J. Dorrestijn, D.T. Crommelin, A. P. Siebesma, and H.J. Jonker, *Stochastic parameterization of shallow cumulus convection estimated from high-resolution model data*, Theor. Comp. Fluid Dyn., 27(1-2), 133–148, 2013.
[5] J. Dorrestijn, D.T. Crommelin, A.P. Siebesma, H.J. Jonker, and C. Jakob, *Stochastic parameterization of convective area fractions with a multicloud model inferred from observational data*, J. Atmos. Sci., 72(2), 854-869, 2015.
[6] W. E, B. Engquist, X. Li, W. Ren, and E. Vanden-Eijnden, *Heterogeneous multiscale methods: a review*, Commun. Comput. Phys., 2(3), 367–450, 2007.
[7] G. Ford and M. Kac, *On the quantum Langevin equation*, J. Stat. Phys., 46(5-6), 803–810, 1987.
[8] J.S. Frederiksen and A.G. Davies, *Eddy viscosity and stochastic backscatter parameterizations on the sphere for atmospheric circulation models*, J. Atmos. Sci., 54(20), 2475–2492, 1997.
[9] I. Grooms and A.J. Majda, *Efficient stochastic superparameterization for geophysical turbulence*, P. Natl. Acad. Sci. USA, 110(12), 4464–4469, 2013.
[10] I.G. Kevrekidis and G. Samaey, *Equation-free multiscale computation: Algorithms and applications*, Annu. Rev. Phys. Chem., 60, 321–344, 2009.
[11] A.J. Majda, I. Timofeyev, and E. Vanden-Eijnden, *Systematic strategies for stochastic mode reduction in climate*, J. Atmos. Sci., 60(14), 1705–1722, 2003.
[12] K. Nimsaila and I. Timofeyev, *Markov chain stochastic parametrizations of essential variables*, Multiscale Model. Sim., 8(5), 2079–2096, 2010.
[13] G. Pavliotis and A. Stuart, *Parameter estimation for multiscale diffusions*, J. Stat. Phys., 127(4), 741–781, 2007.
[14] R. Plant and G.C. Craig, *A stochastic parameterization for deep convection based on equilibrium statistics*, J. Atmos. Sci., 65(1), 87–105, 2008.
[15] P. Porta Mana and L. Zanna, *Toward a stochastic parameterization of ocean mesoscale eddies*, Ocean Model., 79, 1–20, 2014.
[16] A. Stuart and J. Warren, *Analysis and experiments for a computational model of a heat bath*, J. Stat. Phys., 97(3-4), 687–723, 1999.
[17] C.Y. Tang and S.X. Chen, *Parameter estimation and bias correction for diffusion processes*, J. Econometrics, 149(1), 65–81, 2009.
[18] D. Wilks, *Superparameterization and statistical emulation in the Lorenz '96 system*, Q.J. Roy. Meteor. Soc., 138(666), 1379–1387, 2012.
[19] R. Zwanzig, *Nonlinear generalized Langevin equations*, J. Stat. Phys., 9(3), 215–220, 1973.