Data-based parameter estimation of generalized multidimensional Langevin processes

Illia Horenko, Carsten Hartmann, and Christof Schütte
Institut für Mathematik II, Freie Universität Berlin, Arnimallee 2-6, 14195 Berlin, Germany

Frank Noe
IWR, Universität Heidelberg, Im Neuenheimer Feld 368, 69120 Heidelberg, Germany
(Received 12 April 2007; published 19 July 2007)

The generalized Langevin equation is useful for modeling a wide range of physical processes. Unfortunately its parameters, especially the memory function, are difficult to determine for nontrivial processes. We establish relations between a time-discrete generalized Langevin model and discrete multivariate autoregressive (AR) or autoregressive moving average models (ARMA). This allows a wide range of discrete linear methods known from time series analysis to be applied. In particular, the determination of the memory function via the order of the respective AR or ARMA model is addressed. The method is illustrated on a one-dimensional test system and subsequently applied to the molecular dynamics time series of a biomolecule that exhibits an interesting relationship between the solvent method used, the respective molecular conformation, and the depth of the memory.

DOI: 10.1103/PhysRevE.76.016706 PACS number(s): 02.70.Ns, 05.70.Ln

I. INTRODUCTION

The generalized Langevin equation is a useful tool for modeling a wide range of processes, including solid-state, fluid, molecular, and financial dynamics [1–6]. For these high-dimensional systems, the Langevin approach is especially attractive, as it focuses on modeling the dynamics in terms of the few "most important" or "essential" degrees of freedom that carry most of the dynamical information of the observed process.

The formal relationship between the deterministic (and possibly high-dimensional) equations of motion is given by the method of Mori [7] and Zwanzig [8] originating from the field of nonequilibrium thermodynamics. The key element of this procedure is a projection operator, which projects the full set of equations onto the subspace of the essential variables (which needs to be determined in advance). The resulting model is a *generalized Langevin equation* in terms of the essential variables where the effect of the unresolved degrees of freedom (i.e., the variables orthogonal to the essential subspace) is incorporated in terms of a heat bath, involving memory and noise. Further details of the Mori-Zwanzig approach can be found in [9–12]. Unfortunately, the parameters in the generalized Langevin equation are difficult to estimate for nontrivial processes, particularly in multiple dimensions.

In the field of time series analysis, a straightforward and general approach to describing memory effects is given by the (S)ARIMA [(seasonal) autoregressive integrable models with moving average] framework [13,14]; in case of multidimensional applications, these are also frequently called *multivariate autoregressive* (MVAR) models [15]. So-called autoregressive moving average (ARMA) models represent the class of *time-discrete* (S)ARIMA models that involve memory and *additive* noise. Robust methods for parametrizing ARMA models by fitting the observed data with a discrete time stochastic difference scheme are available. The feasibility of (S)ARIMA-type models for the qualitative description of (low-dimensional) molecular dynamics data has been demonstrated in [16].

In the present paper, we establish relations among the time-discrete generalized Langevin equation and time-discrete multivariate autoregressive (AR) or ARMA model, respectively. This allows for applying a wide range of discrete linear methods that originate from time-series analysis to the parameter estimation problem of the generalized Langevin equation [13,17]. In particular, we focus on determining the order of the respective AR or ARMA model, indicating the depth of memory in the given data. The method is illustrated by means of a simple one-dimensional test system and is subsequently applied to the simulated dynamics of the 8-alanine peptide. The relationship between solvent model (explicit or implicit water), molecular conformation (α-helix or β-hairpin), and the depth of the memory is investigated.

II. GENERALIZED LANGEVIN EQUATION

Let $x_t \in \mathbf{R}^n$ be a time-dependent vector of some (essential) degrees of freedom and let $v_t \in \mathbf{R}^n$ be the corresponding velocity vector. For many physical processes the dynamics in these variables can be described by a generalized Langevin equation of the form

$$\dot{x}_t = v_t,$$

$$M\dot{v}_t = -\nabla U(x_t) - \int_0^t \gamma(t-s)v_s ds + F_t, \qquad (1)$$

where $M \in \mathbf{R}^{n\times n}$ is the symmetric, positive-definite mass matrix, $\gamma \in \mathbf{R}^{n\times n}$ denotes the positive semidefinite memory

*horenko@math.fu-berlin.de
†frank.noe@iwr.uni-heidelberg.de
‡chartman@math.fu-berlin.de
§schuette@math.fu-berlin.de

 ©2007 The American Physical Society

kernel, and F_t is a zero-mean stochastic process in $\mathbf{R}^n$. The potential energy $U:\mathbf{R}^n \to \mathbf{R}$ is bounded from below. It is important to note that in the thermodynamic sense this is an open system since only an (essential) subspace of the full phase space is modeled. Consequently we cannot assume that the system is in thermodynamic equilibrium, and hence no fluctuation-dissipation relation between the memory kernel and the noise process is imposed from the outset. We understand the generalized Langevin equation in the present form (1) as a phenomenological model, the derivation of which involves quite a number of approximations that we do not want to discuss here; the interested reader is referred to the recent textbook [18]. Although appealing, the generalized Langevin equation (1) is difficult to parametrize without further restrictions. Throughout this paper we set in force the following model assumptions.

(i) The memory kernel is a piecewise constant function.

(ii) The stochastic process is white noise with covariance matrix $\sigma\sigma^T \in \mathbf{R}^{n\times n}$—i.e., $F_t = \sigma \dot{W}_t$, where $W_t \in \mathbf{R}^n$ denotes standard Brownian motion.

(iii) The potential U can be (locally) approximated by a harmonic potential:

$$U(x) = \frac{1}{2} x^T D x,$$

with $D \in \mathbf{R}^{n\times n}$ being the corresponding stiffness matrix. Note that the equilibrium position of the harmonic potential is set to zero. This is done for convenience of notation and can be generalized.

Assumption (i) is pragmatic, taking account of the fact that any observation data upon which a parametrization is built are discrete. Thus for a sufficiently small time step between the data points, every memory kernel can be replaced by a suitable step function. Assumption (ii) is supported by the central limit theorem for weakly dependent Gaussian processes [19] and the physical idea that the bath fluctuations stem from weakly coupled nonlinear oscillations of those molecular degrees of freedom not incorporated into the Langevin model (1). Such oscillations are typically fast as compared to the characteristic time scale of the essential dynamics which explains the absence of correlations in the noise. Assumption (iii) is physically reasonable as long as the trajectory resides within a localized region of state space, such as a single-molecular conformation. (In principle, even more general processes can be described in terms of a set of locally harmonic models; cf. [20] and the discussion in Sec. VI below.)

In order to determine the parameters in the generalized Langevin equation (1), two possible approaches are pursued.

From generalized Langevin to AR models. On condition that the assumptions above hold true, we can recast the generalized Langevin equation in the standard form of a linear autoregressive model AR(q). The optimal parameters of the AR(q) model are computed using standard estimators, thus obtaining information about the depth of the memory in the input time series. As the numerical effort of the parameter estimation scales as $\mathcal{O}(d^6)$, where $d=2n$ is the phase space dimension, the approach is limited to considerably small problems.

From generalized Langevin to ARMA models. Since the parameter estimation procedure for AR models scales very unfavorably with the phase-space dimension d, it is important to restrict the model to only a few essential variables that span a linear subspace of the configuration space. These are called the resolved variables. If the remaining (unresolved) variables have quickly decaying autocorrelations and Gaussian-like probability distributions, we can replace them by suitable Gaussian processes, which then leads to a low-dimensional ARMA(q,p) model. In addition to the AR(q) memory, the ARMA(q,p) model exhibits additional memory stemming from the interplay between resolved and unresolved degrees of freedom.

III. FROM GENERALIZED LANGEVIN TO AR MODELS

Suppose that the modeling assumptions (i)–(iii) hold true. The generalized Langevin equation (1) then reads

$$\dot{x}_t = v_t,$$

$$M\dot{v}_t = -Dx_t - \sum_{i=0}^{q} \gamma_i v_{t-i\tau} + \sigma \dot{W}_t, \tag{2}$$

where $\gamma_i = \gamma(t-i\tau)$. The memory depth $q \in \mathbf{N}$ is determined by the memory kernel's support backward in time. Introducing the shorthand notation $Q=(x,v)$ the generalized Langevin equation (2) can be written as

$$\dot{Q}_t = AQ_t + \sum_{i=1}^{q} \Gamma_i Q_{t-i\tau} + \Sigma \dot{B}_t, \tag{3}$$

with $B_t = (0, W_t)^T \in \mathbf{R}^{2n}$ and the matrices

$$A = \begin{pmatrix} 0 & I \\ -M^{-1}D & -M^{-1}\gamma_0 \end{pmatrix}, \quad \Gamma_i = \begin{pmatrix} 0 & 0 \\ 0 & -M^{-1}\gamma_i \end{pmatrix},$$

$$\Sigma = \begin{pmatrix} 0 & 0 \\ 0 & M^{-1}\sigma \end{pmatrix}.$$

Assuming τ is small, we can apply the Euler-Maruyama discretization of Eq. (3). The resulting numerical scheme can be written in the form

$$Q_{t+\tau} = \sum_{i=0}^{q} \alpha_i(\tau) Q_{t-i\tau} + \beta(\tau)\epsilon_t, \tag{4}$$

with the abbreviations

$$\alpha_0(\tau) = \begin{pmatrix} I & \tau I \\ -\tau M^{-1}D & I - \tau M^{-1}\gamma_0 \end{pmatrix},$$

$$\alpha_i(\tau) = \begin{pmatrix} 0 & 0 \\ 0 & -\tau M^{-1}\gamma_i \end{pmatrix} \text{ (for } i > 0),$$

$$\beta(\tau) = \begin{pmatrix} 0 & 0 \\ 0 & \sqrt{\tau} M^{-1} \sigma \end{pmatrix},$$

and $\epsilon_t = (0, \xi_t)^T$, where $\xi_t \sim \mathcal{N}(0, I)$ is a Gaussian process with zero mean and unit variance. Equation (4) has the form of the standard linear autoregressive model AR(q) [13]. The coefficients $\alpha_i(\tau)$ are called *partial autocorrelation coefficients* (PACCs). Multiplying both sides of Eq. (4) with $Q^T_{t-j\tau}$, $j=0,\ldots,q$, from the right yields upon taking the expectation value

$$\mathbf{E} Q_{t+\tau} Q^T_{t-j\tau} = \sum_{i=0}^{q} \alpha_i(\tau) \mathbf{E} Q_{t-i\tau} Q^T_{t-j\tau} + \beta(\tau) \mathbf{E} \epsilon_t Q^T_{t-j\tau}.$$

Since the realizations of the Gaussian white noise process are independent of the values of $Q_{t-j\tau}$, $j=0,\ldots,q$, the rightmost term in the last equation vanishes. If the process Q_t is assumed to be weakly stationary[1] and the covariance matrix of the process is invertible, then it follows that the cross-covariance matrix is symmetric with respect to time shifts—i.e., $\mathbf{E} Q_t Q^T_{t-j\tau} = \mathbf{E} Q_t Q^T_{t+j\tau}$. Thus we obtain

$$\mathbf{E} Q_t Q^T_{t-(j+1)\tau} = \sum_{i=0}^{q} \alpha_i(\tau) \mathbf{E} Q_t Q^T_{t-(i-j)\tau}. \tag{5}$$

The autocorrelation matrix

$$c_j(\tau) = \mathbf{E} Q_t Q^T_{t-j\tau} (\mathbf{E} Q_t Q_t^T)^{-1} \tag{6}$$

can be easily calculated from the observed data sequence Q_t. By multiplying both sides of Eq. (5) by the inverse of the covariance matrix $\mathbf{E} Q_t Q_t^T$ from the right and substituting Eq. (6), we obtain $(q+1)$ linear equations with $(q+1)$ unknown matrices, known as the *Yule-Walker system*,

$$c_{j+1}(\tau) = \sum_{i=0}^{q} \alpha_i(\tau) c_{i-j}(\tau), \quad i,j = 0, \ldots, q, \tag{7}$$

from which the $(q+1)$ matrix-valued partial autocorrelation coefficients $\alpha_i(\tau)$ can be calculated [13,17]. For a given d-dimensional observation sequence Q_t, the identification procedure consists of two basic steps: selection of the order q and then estimation of the respective AR(q) parameters. The order q can be chosen according to the Box-Jenkins scheme [13,17]: One calculates the PACCs for sufficiently large q from the autocorrelation matrices [14]. The actual order of the AR model is then determined by the number of PACCs that lie outside the confidence interval $[-1/\sqrt{T}, 1/\sqrt{T}]$, where $T = N\tau$ is the total length of the observed time series Q_t. The order q of the AR model indicates the "depth" of the memory in the data.

From a numerical point of view, the calculation of PACCs becomes expensive with a growing number of dimensions $d=2n$: The Yule-Walker equation has d^2 unknowns. Roughly speaking, its solution requires $\mathcal{O}(d^6)$ operations. If the resulting matrix is sparse with $\mathcal{O}(d)$ entries or is diagonally dominant, then the number of operations becomes $\mathcal{O}(d^2 \ln(d))$. This large numerical effort underpins the need of dimension reduction before estimating the AR(q) parameters, as will be discussed next.

[1]This means that both $\mathbf{E} Q_t = \mathbf{E} Q_s$ are independent of time and $\mathbf{E} Q_t Q^T_{t-j\tau} = \mathbf{E} Q_s Q^T_{s-j\tau}$ holds true for all $s,t \in [0,T]$. In cases where weak stationarity of the data is not obvious, this may be checked with the *unit-root test* [17].

IV. FROM GENERALIZED LANGEVIN TO ARMA MODELS

Let $S \subset \mathbf{R}^n$ be an affine m-dimensional linear configuration subspace. S is assumed to represent the essential dynamics of the system in the sense that all those orthogonal to S degrees of freedom have Gaussian-like distributions with quickly decaying correlations. In this case the system's motion is approximated by the motion in S, whereas the fluctuations in the orthogonal complement $S^\perp = \mathbf{R}^n \backslash S$ can be considered as noise.

We introduce two orthogonal projections Π and $\Pi^\perp = I - \Pi$: Π projects onto the subspace S—i.e., $\Pi x \in S \subset \mathbf{R}^n$—whereas $\Pi^\perp$ projects onto its orthogonal complement $S^\perp$. Without loss of generality, we may define $\Pi = PP^T$ and $\Pi^\perp = RR^T$ with orthogonal matrices $P \in \mathbf{R}^{n \times m}$ and $R \in \mathbf{R}^{n \times (n-m)}$, respectively. By setting $y = P^T x$ and $z = R^T x$, local coordinates on the two subspaces S and $S^\perp$ are obtained. Accordingly, $r = P^T v$ and $s = R^T v$ define local coordinates on the corresponding tangent spaces. In these coordinates, the equations for the dynamics in the resolved subspace read

$$\dot{y}_t = r_t,$$

$$\dot{r}_t = -C_1 y_t - C_2 z_t - \sum_{i=0}^{q} (K_i r_{t-i\tau} + L_i s_{t-i\tau}) + A \dot{W}_t, \tag{8}$$

with the matrices

$$C_1 = P^T M^{-1} D P, \quad C_2 = P^T M^{-1} D R,$$

$$K_i = P^T M^{-1} \gamma_i P, \quad L_i = P^T M^{-1} \gamma_i R, \quad A = P^T M^{-1} \sigma.$$

Clearly Eq. (8) is not closed as it still depends on the unresolved variables. We seek an effective equation for the resolved modes by replacing the unresolved variables by appropriate stochastic processes. By assumption, the unresolved modes are fast with Gaussian distributions. Hence, a systematic closure strategy consists in replacing the unresolved modes in Eqs. (8) by suitable δ-correlated Gaussian processes. This results in

$$\dot{y}_t = r_t,$$

$$\dot{r}_t = -C_1 y_t - \sum_{i=0}^{q} (K_i r_{t-i\tau} + L_i \xi_{t-i\tau}) - C_2 \zeta_t + a \dot{W}_t. \tag{9}$$

Here, ξ_t and ζ_t are stationary Gaussian processes with zero mean satisfying

$$\mathbf{E} \xi_t \xi_s^T = \mathrm{cov}(R^T v) \delta(t-s), \quad \mathbf{E} \zeta_t \zeta_s^T = \mathrm{cov}(R^T q) \delta(t-s).$$

Following the procedure explained in the previous section we discretize the last equation using an Euler-Maruyama

scheme. Defining the state vector $U=(P^T x, P^T v) \in \mathbf{R}^{2m}$, the following discrete iteration is obtained:

$$U_{t+\tau} = \sum_{i=0}^{q} \alpha_i(\tau) U_{t-i\tau} + \sum_{j=0}^{p} \beta_j \epsilon_{t-j\tau} \tag{10}$$

for appropriately defined coefficients $\alpha_i, \beta_j \in \mathbf{R}^{2m\times 2m}$. If the process ϵ_τ is standard white noise, the process (10) is an instance of an ARMA(q,p) model. The order q defines the depth of the "internal" memory of the resolved system itself, whereas p indicates the "external" memory—i.e., the memory resulting from the coupling between the resolved and unresolved modes.

The optimal orders (q,p) for a given time series U_t are obtained using *Akaike's information-corrected criterion* [17]. It proceeds by evaluating the prediction error of a given ARMA(q,p) model in reproducing the time series U_t. After finding the optimal orders (q,p), we employ the *innovation algorithm* to estimate the model parameters [17].

V. ILLUSTRATION OF A LINEAR MARKOVIAN LANGEVIN MODEL

In order to demonstrate the usefulness of the concepts described above and to show how the memory depth determination based on the PACCs works in practice, we consider the Markovian limit of the generalized Langevin equation (1), which is obtained in case the memory kernel has the form of a Dirac function. [Alternatively, one could set $\gamma_i = 0$ for $i>0$ in Eq. (2).] The following one-dimensional linear Langevin model is examined:

$$\dot{x}_t = v_t,$$

$$\dot{v}_t = -Dx_t - \gamma v_t + F_t, \tag{11}$$

with scalar positions and velocities x_t and v_t and scalar random force F_t, which is assumed to be white noise. The autocorrelation function can be calculated analytically (e.g., see [21]):

$$c_j(\tau) = \exp(j\tau F), \quad F = \begin{pmatrix} 0 & 1 \\ -D & -\gamma \end{pmatrix}. \tag{12}$$

The decay behavior of this autocorrelation matrix for $D=1.0$ and two different values of the friction parameter γ is shown in Fig. 1. If the friction γ is small, both the position autocorrelation $c_j^{11}(\tau)$ and the velocity autocorrelation $c_j^{22}(\tau)$ decay equally slowly. If the friction is large, the velocity autocorrelation $c_j^{22}(\tau)$ decays much faster than the position autocorrelation $c_j^{11}(\tau)$. In any event, however, the autocorrelation matrix gives no clue regarding the memory depth of the underlying system (which is Markovian by construction).

An unambiguous way to calculate the memory of system (11) is provided by the partial autocorrelation coefficients $\alpha_i(\tau)$, which can be calculated by solving the Yule-Walker system (7). As Fig. 2 clearly indicates, all PACCs α_i are zero except for $i=0$—independently of the friction γ. Hence, the PACCs correctly reveal the Markovian character of the data. Moreover, this property is independent of the discretization lag τ as can be readily shown.

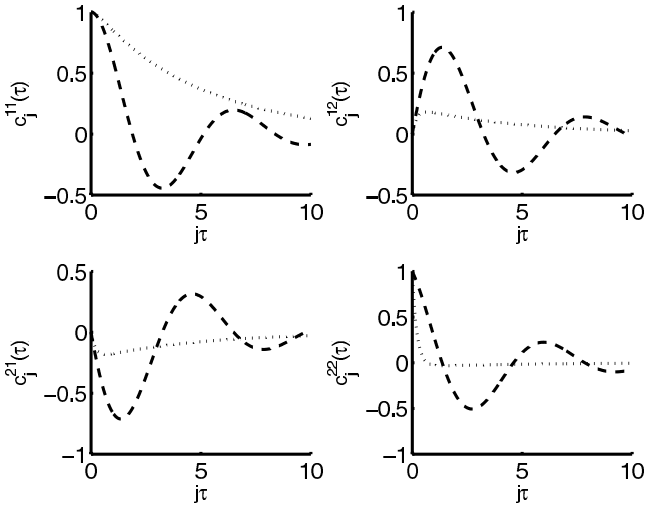

FIG. 1. Autocorrelation matrix $c_j(\tau)$ of the (Markovian) Langevin equation (11) for $D=1$ and two different values of the friction coefficient: $\gamma=5$ (dotted lines) and $\gamma=0.5$ (dashed lines). The rate of decay depends on the friction parameter but has nothing to do with possible memory in the data.

Note that the picture may change if the observation data are incomplete. For example, if we pick out only the velocity autocorrelation component to calculate the corresponding scalar partial autocorrelation, then there are indeed nonvanishing PACCs for $i>0$, which could be misinterpreted as arising from non-Markovian dynamics (compare Fig. 3). However, the velocity component alone is a non-Markovian process (since it depends on the current position) and it is also nonstationary, for it depends on the neglected position component. Hence also, the velocity autocorrelation function is explicitly time dependent by virtue of the positions x_t, which prohibits the direct solution of the Yule-Walker system.

We emphasize that stationarity is very important for the applicability of the concepts described above. Nonstationar-

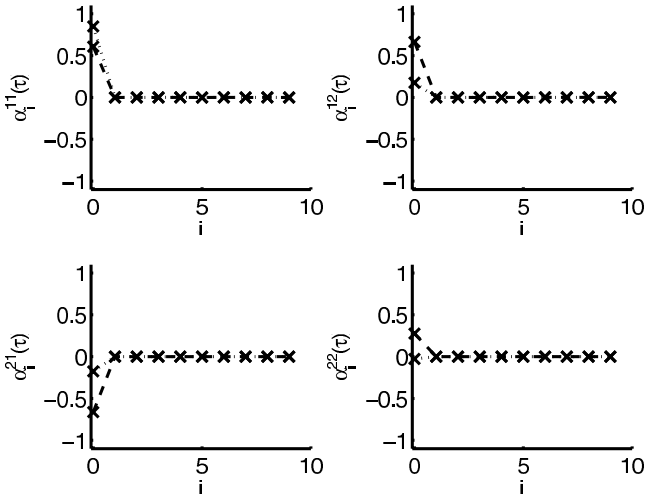

FIG. 2. Partial autocorrelation matrix $\alpha_i(\tau)$ of the (Markovian) Langevin equation (11). For $i>0$ all calculated partial autocorrelation coefficients α_i are zero, correctly indicating the systems' Markovian property. (Line styling as in Fig. 1 above.)

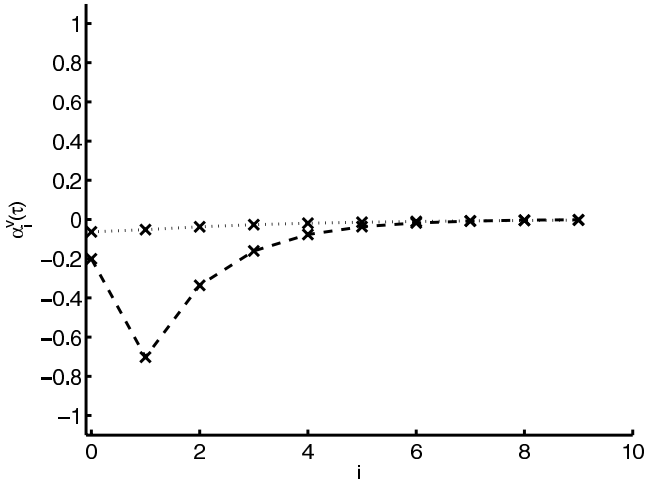

FIG. 3. Partial autocorrelation coefficient $\alpha_i^{vel}(\tau)$ resulting from the solution of the Yule-Walker system with the velocity autocorrelation $c_j^{22}(\tau)$ only at different time lags $i\tau$ (τ=1.0). For $i>0$ the calculated partial autocorrelation coefficients α_i are nonzero which may indicate a non-Markovian dynamics.

ity may arise, for example, if the underlying system exhibits metastability, such that expectation values do not converge due to finite sampling. In such (frequently occurring) cases it is possible to *stationarize* the analyzed data—e.g., by identifying the metastable states and separating the time series into corresponding subseries as will be demonstrated next.

VI. DYNAMICS OF THE 8-ALANINE PEPTIDE IN WATER: RELATIONSHIP BETWEEN SOLVENT MODEL, CONFORMATION, AND MEMORY

Molecular dynamics simulation. Molecular dynamics simulations were performed for the 8-alanine peptide Ala_8 with zwitterionic termini at temperature T=300 K using the CHARMM force field [22]. The parameter set number 19 was used to model intramolecular interactions. Four sets of simulations were carried out, in α-helical and β-hairpin configurations, using implicit and explicit solvent models. The Verlet integration time step was 2 fs, keeping the hydrogen bond lengths constant.

For the implicit solvent simulations, we used the ACE2 method to model electrostatic effect of the solvent [23]. A switch function was used to fade out the nonbonded interaction energies between 8 and 12 Å. Both starting conformations α and β were minimized up to 10^{-3} kcal/(mol Å) of the root mean square of the energy gradient. The energy minimization proceeded in three subsequent stages: 1000 steps of steepest descent minimization, 1000 steps of conjugate gradient minimization, and finally 4000 steps of Newton-Raphson minimization. The system was then heated to the simulation temperature of T=300 K during 20 ps. A local equilibration (to relax the system within its local conformation) was performed over a period of 60 ps, followed by an unconstrained production run of 1 ns. The temperature was kept constant using ordinary temperature rescaling. The

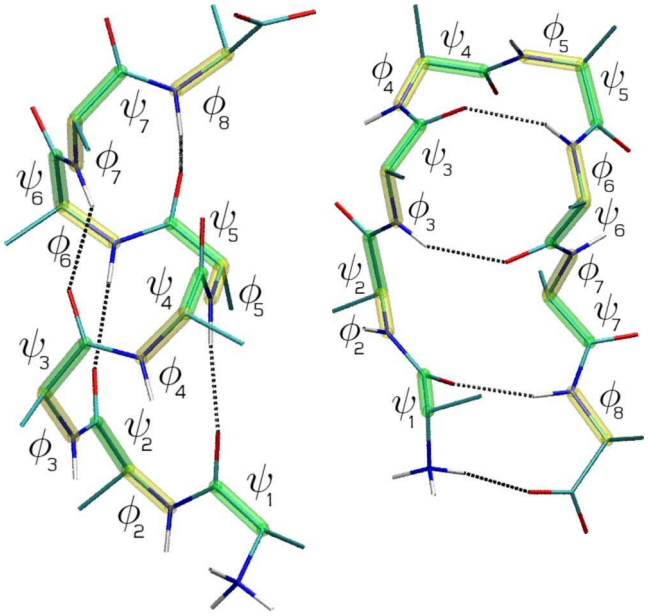

FIG. 4. (Color online) Main configurations of Ala_8 in water: α-helix (left) and β-hairpin (right).

solvent viscosity was not incorporated in any way. Given these simulation conditions, Ala_8 exhibits two metastable states, the mean structures of which are shown in Fig. 4.

In the explicit solvent simulations, the peptides were embedded in a truncated octahedral box of TIP3 water with a diameter of 30 Å, yielding 386 water molecules for the α and 381 water molecules for the β conformation. A switch function was used to fade out the nonbonded interaction energies between 10 and 13 Å. The solvated boxes were energy-minimized up to 10^{-2} kcal/(mol Å) of the root mean square of the energy gradient, proceeding in two subsequent stages: 5000 steps of steepest descent minimization and 5000 steps of Newton-Raphson minimization. It was then heated to the simulation temperature of T=300 K during 20 ps using weak positional harmonic constraints [0.1 kcal/(mol Å)] to keep the peptide atoms in place. A local equilibration (to relax the system within its local conformation) was conducted for 120 ps, followed by an unconstrained production run of 1 ns. Both pressure and temperature were kept constant using the Berendsen thermo-barostat.

For the analysis, both coordinates and velocities were saved in each integration step (every 2 fs). The actual observation data was generated in the following way: Let $x_t=f(q_t)$ be the observed variable (e.g., torsion angles $f_i:\mathbf{R}^{12}\rightarrow S^1$), where q_t denotes the trajectory of the full molecule in Cartesian space. The corresponding velocity vector $v_t=\dot{x}_t$ was then calculated by numerical Euler differentiation,

$$v_t=\frac{f(q_t+h\dot{q}_t)-f(q_t-h\dot{q}_t)}{2h},\quad h=0.001\text{ fs}.$$

Parameter estimation: AR model. To determine the order of the AR(q) model in terms of the backbone angles $f=(\psi_1,\phi_1,\ldots,\psi_7,\phi_8)$, we first calculate the time-dependent autocorrelation matrices $c_j(\tau)\in\mathbf{R}^{28\times 28}$ (τ=2 fs) of the

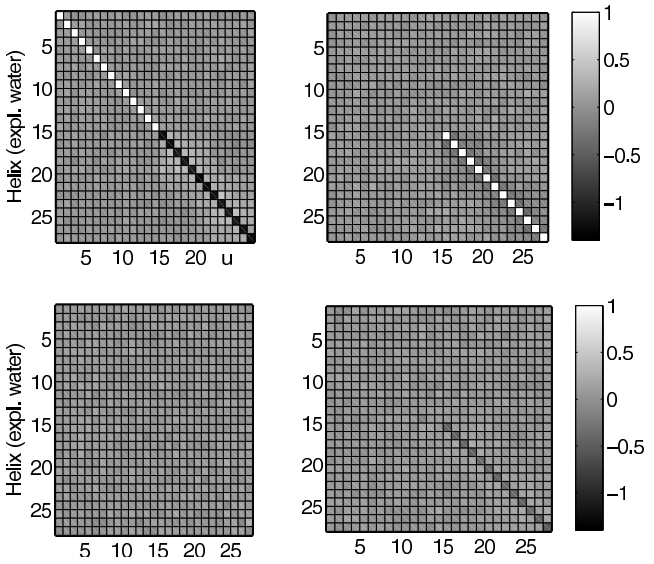

FIG. 5. Parameters α_0, α_1 (upper row) and α_2, α_3 (lower row) of the discrete AR model (4) for the helical conformation with explicit solvent.

angles and its velocities (here $j=1,2,\ldots$ is the discrete time index); the torsion angles are numbered along the backbone as depicted in Fig. 4. The Yule-Walker method [14,17] is employed to calculate the AR(q) parameters from the autocorrelation matrices (see Fig. 5). As is apparent from Fig. 6, the PACCs decay rather quickly within a few femtoseconds, whereas the eigenvalues of the autocorrelation matrices do not decay within 100 fs. For the explicit water model the PACCs decay slower than in implicit solvent; i.e., the explicit solvent introduces additional memory into the system.

The discrete memory kernel γ_i in the generalized Langevin equation (2) is then obtained from the PACC. Figure 7 shows the estimates of $M^{-1}\gamma_i, i=0,\ldots,3$, for different molecular configurations and solvent models. Note that the mass matrix cannot be estimated explicitly unless the system is assumed to be in thermodynamic equilibrium (fluctuation-dissipation relation) [21]. It turns out that the parameter matrices of the two conformations are quite different: the matrices corresponding to the α-helical configuration have a bandlike structure, whereas the β-hairpin matrices appear to have two blocks corresponding to the two β-sheets of the hairpin. These differences are most pronounced in the Mar-

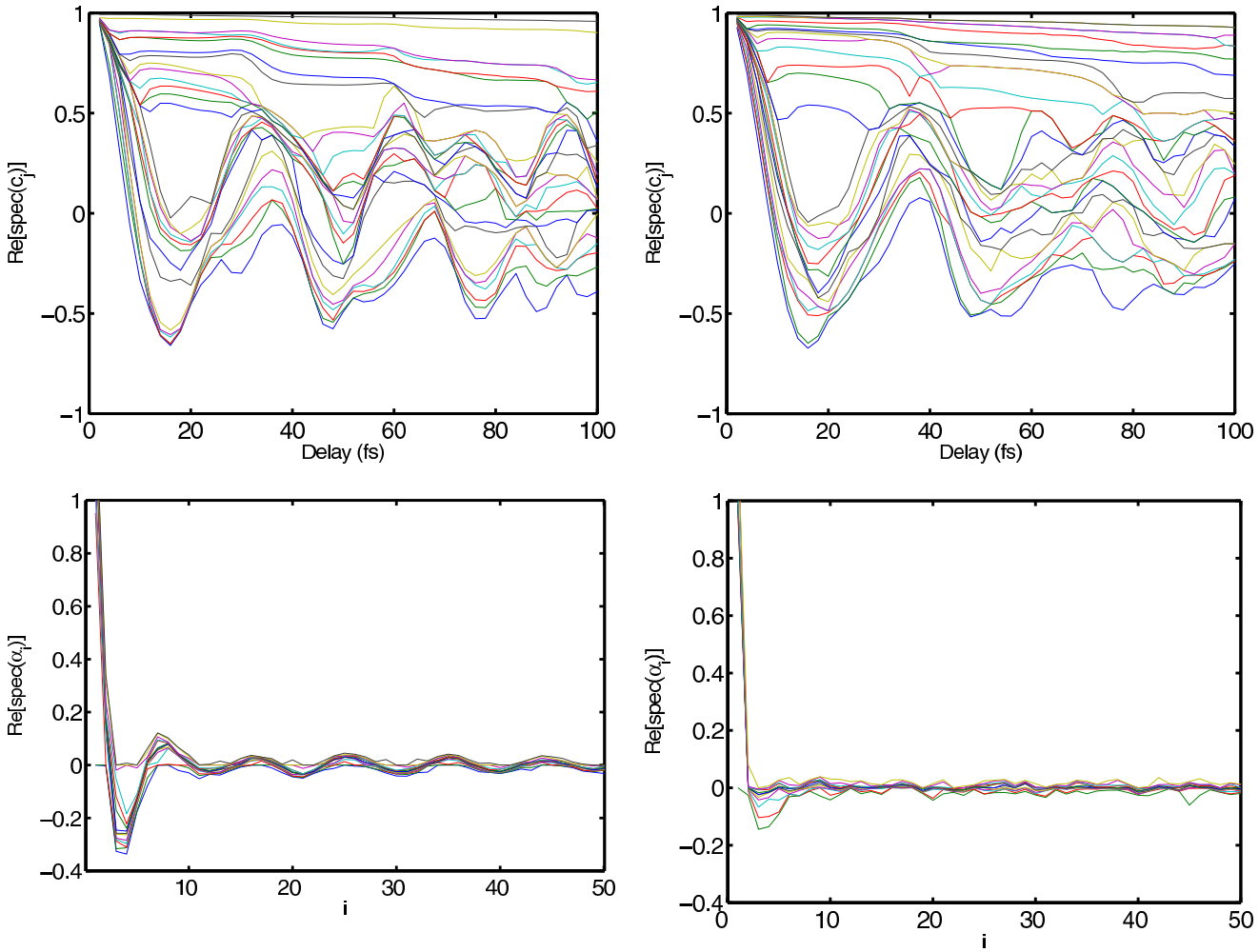

FIG. 6. (Color online) Simulation of the Ala_8-peptide in the α-helical conformation. Left panel: real part of the eigenvalues of the autocorrelation matrices $c_j(\tau)$ for explicit (left) and implicit (right) water models. Right panel: real part of the eigenvalues of the partial autocorrelations $\alpha_i(\tau)$. The x axis in all cases spans the same time interval between 0.0 and 100.0 fs.

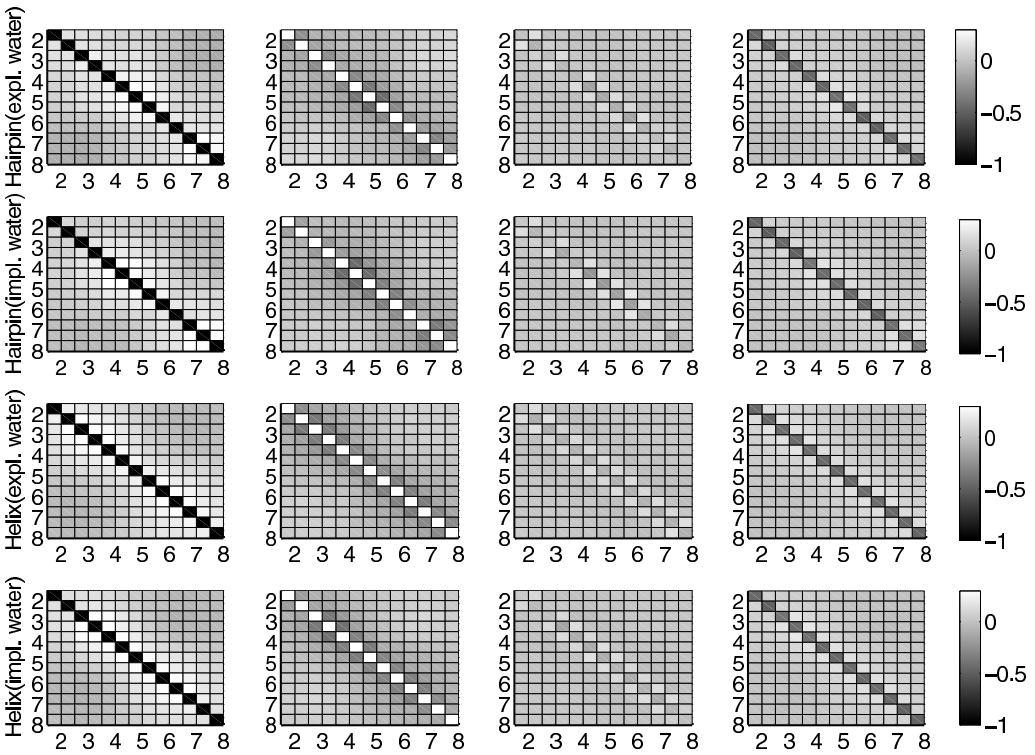

FIG. 7. Memory kernel matrices $M^{-1}\gamma_i$,i=0,…,3, as estimated from the time series of the torsion angles and their velocities (cf. Fig. 4).

kovian part of the kernel (i.e., for i=0), and they become weaker as i increases.

Comparison of the matrices for different solvent models leads to the following observations: (a) the occupancy pattern of the memory matrix is more or less independent of the solvent model, and (b) the absolute values of the off-diagonal entries for the implicit water model are smaller than for the explicit solvent that results in an overall lower friction in the implicit solvent model. This is consistent with the fact that the expected exit times for α and β conformations are lower in the implicit than in the explicit solvent simulations (data not shown).

Parameter estimation: ARMA model. The next step is to estimate the order parameters for the ARMA(q,p) model, given a series of torsion angles and torsion velocities. To identify the essential torsion subspace the method of optimal persistence patterns (OPPs) is employed since it allows a data-based separation of fast and slow modes in multidimensional time series [24]. By maximizing the functional

$$\mathcal{L}(P) = \sum_{i=1}^{\infty} \mathrm{tr}[P^T c_i(\tau) P], \tag{13}$$

OPPs find an m-dimensional affine subspace S that is defined by an orthogonal matrix $P \in \mathbf{R}^{n\times m}$ with $PP^T x \in S$ for all $x \in \mathbf{R}^n$. The subspace is characterized by the slowest possible decay of autocorrelations, where the projected autocorrelation matrices $P^T c_i P \in \mathbf{R}^{m\times m}$ are clearly the autocorrelation matrices of the projected data $y = P^T x$ (its trace is the sum of the autocorrelation functions of the projected data). For details on the maximization of the functional (13), we refer to the article [24].

Application of the OPP method to the Ala_8 torsion angle data (without velocities) results in the normalized integrated autocorrelation spectra shown in Fig. 8. For the helical conformation and for the hairpin configuration with implicit solvent one clearly pronounced slow mode is observed, whereas there is no clear time scale separation in the explicit solvent hairpin simulation. The OPP coordinates are shown in Fig. 9. If we take the first $m \geqslant 1$ modes as essential slow coordi-

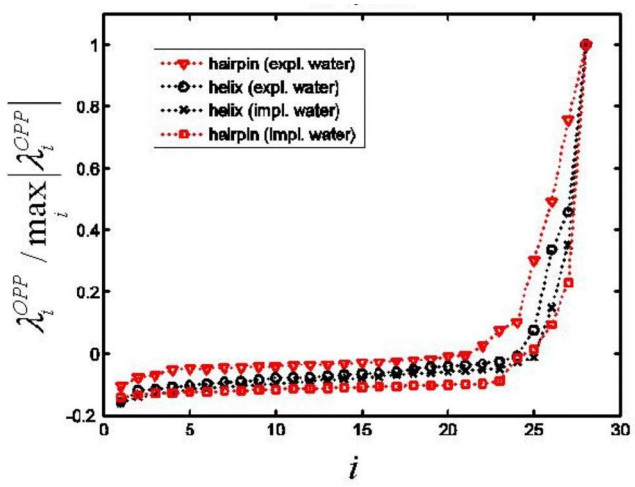

FIG. 8. (Color online) Integrated eigenvalues of $\Sigma_{i=1}^{\infty} c_i(\tau)$ for different conformations and solvent models. The values are normalized, such that the rightmost value 1 corresponds to the slowest mode in the system.

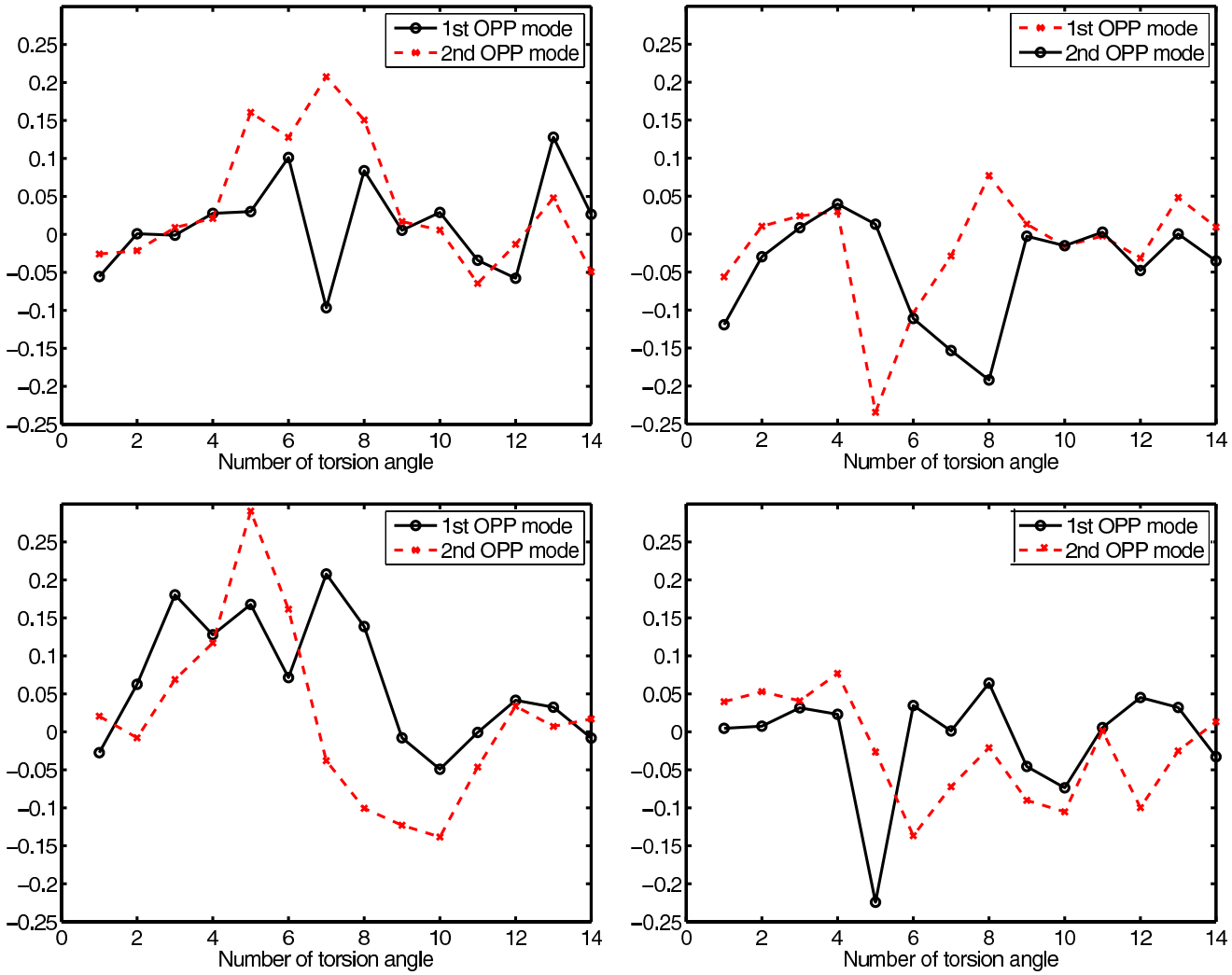

FIG. 9. (Color online) Dominant OPP eigenvectors (columns of the matrix P for $m=2$).

nates, then the probability distribution of $n-m$ remaining unresolved fast modes is essentially Gaussian. Thus, assuming that the correlations of the unresolved modes decay sufficiently fast, the derivation of the reduced ARMA model (10) is valid, if at least the first dominant torsional OPP mode is resolved.

In the first step, a one-dimensional reduced subspace ($m=1$) is considered. This yields an ARMA(q,p) model with a two-dimensional phase space. The optimal values of q,p are obtained using Akaike's order selection criterion. As apparent from Table I there is no clear dependency of the order q of the autoregressive part (internal memory) on the solvent model. In contrast, the order p of the moving average part (external memory) is consistently higher for the implicit water model. For the sake of illustration Table II shows the optimal values of q,p for $m=2$. The external memory with explicit water is dramatically increased as compared to $m=1$. Furthermore, the internal memory of the α-helix is more than twice as long as the internal memory of the β-hairpin, which may be due to the fact that the α conformation allocates a smaller region of torsion angle space than the β conformation and is thus more "coherent." (The memory depth in physical time is obtained upon multiplying the numbers q and p by the discretization step size $\tau=2$ fs.)

TABLE I. Optimal order parameters p,q for $m=1$.

ARMA(q,p)	Implicit water	Explicit water
α-helix	$q=5,p=7$	$q=7,p=4$
β-hairpin	$q=2,p=7$	$q=2,p=4$

TABLE II. Optimal order parameters p,q for $m=2$.

ARMA(q,p)	Implicit water	Explicit water
α-helix	$q=5,p=7$	$q=6,p=8$
β-hairpin	$q=2,p=6$	$q=2,p=8$

VII. CONCLUSIONS

The correct parametrization of the *generalized Langevin equation* given multidimensional observation data is a major problem in the modeling of many physical processes. The

main contribution of the present work is that it establishes a connection between the time-discrete form of the generalized Langevin equation and discrete (S)ARIMA models. This allows for the application of a wide range of methods known from the classical time series analysis to the generalized Langevin equation.

The main obstacle in applying the methods to the generalized Langevin equation is the enormous numerical cost of the AR parameter identification procedure [in general $\mathcal{O}(d^6)$, where d is the number of degrees of freedom used in the analysis]. This restricts the applicability of the AR-fitting procedure to low-dimensional cases. Thus, an appropriate dimension reduction is essential in order to efficiently estimate AR models. It is shown how the further dimension reduction of the generalized Langevin model leads to discrete ARMA models with an additional memory term related to unresolved degrees of freedom.

The procedure was illustrated for a one-dimensional model system, proving that the popular idea that (non-Markovian) memory is indicated by the decay of the autocorrelation is misleading. Instead partial autocorrelation coefficients are to be considered. Moreover, we pointed out that an important requirement for the applicability of the parameter estimation procedure is the *weak stationarity* of the observed time series. Only in this case do the Yule-Walker equations (7) hold and can be used to determine the PACCs required for the estimation of the memory function. This means that this property should be checked prior to the parameter estimation procedure.

The practical usefulness of the method was shown by the application to molecular dynamics data of a realistic peptide molecule in different solvents. The results confirm the intuitive idea that for molecular systems the use of explicit solvent considerably increases the memory of the physical model with respect to implicit solvent models. This emphasizes the requirement of incorporating additional friction into implicit solvent simulations to have a physically reasonable model—e.g., by means of a Langevin thermostat. More generally, the results indicate that the actual memory pertaining to the observed physical system may often be much shorter than the decay of the autocorrelation function suggests.

ACKNOWLEDGMENT

This work was supported by the DFG Research Center MATHEON "Mathematics for key technologies."

[1] S. A. Adelman and J. D. Doll, J. Chem. Phys. **64**, 2375 (1976).
[2] J. B. Witkoskie, J. Wu, and J. Cao, J. Chem. Phys. **120**, 5695 (2004).
[3] J. Peinke, A. Kittel, S. Barth, and M. Oberlack, in *Progress in Turbulence*, edited by B. Dubrulle, J.-P. Lavale, and S. Nazarenko (Springer, Berlin, 2005), pp. 77–86.
[4] T. Yamaguchi, T. Matsuoka, and S. Koda, J. Chem. Phys. **122**, 014512 (2005).
[5] M. Takahashi, Financial Eng. Jpn. Markets **3**, 87 (1996).
[6] O. Lange and H. Grubmüller, J. Chem. Phys. **124**, 214903 (2006).
[7] H. Mori, Prog. Theor. Phys. **33**, 423 (1965).
[8] R. W. Zwanzig, J. Stat. Phys. **9**, 215 (1975).
[9] J. T. Hynes and J. M. Deutch, in *Physical Chemistry—An Advanced Treatise*, edited by H. Eyring, D. Henderson, and W. Jost (Academic, New York, 1975), pp. 729—836.
[10] R. W. Zwanzig, in *Systems Far From Equilibrium*, edited by J. Ehlers, K. Hepp, R. Kippenhahn, H. A. Weidenmüller, and J. Zittarz (Springer, Berlin, 1980).
[11] D. J. Evans and G. P. Morriss, *Statistical Mechanics of Nonequilibrium Liquids* (Academic, London 1990).
[12] D. Givon, O. H. Hald, and R. Kupferman, Isr. J. Math. **145**, 221 (2005).
[13] G. Box and G. Jenkins, *Time Series Analysis, Forecasting, and Control* (Holden-Day, San Francisco, 1976).
[14] S. M. Kay, *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Proceedings* (IEEE, New York, 2003), Vol. 3, pp. 289–292.
[15] P. Premakanthan and W. B. Mikhael (unpublished).
[16] A. Gorecki, J. Trylska, and B. Lesyng, Europhys. Lett. **75**, 3 (2006).
[17] P. J. Brockwell and R. A. Davis, *Introduction to Time Series and Forecasting* (Springer, Berlin 2002).
[18] A. J. Chorin and O. H. Hald, *Stochastic Tools in Mathematics and Science* (Springer, New York, 2006).
[19] A. M. Stuart and R. Kupferman, Physica D **199**, 279 (2004).
[20] I. Horenko, E. Dittmer, A. Fischer, and Ch. Schütte, Multiscale Model. Simul. **5**, 802 (2006).
[21] I. Horenko and C. Schütte (unpublished).
[22] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus, J. Comput. Chem. **4**, 187 (1983).
[23] M. Schaefer and M. Karplus, J. Chem. Phys. **100**, 1578 (1996).
[24] T. DelSole, J. Atmos. Sci. **58**, 1341 (2001).