Discrete approach to stochastic parametrization and dimension reduction in nonlinear dynamics

Alexandre J. Chorin[a,b,1] **and Fei Lu**[a,b]

[a]Department of Mathematics, University of California, Berkeley, 94720; and [b]Mathematics Group, Lawrence Berkeley National Laboratory, Berkeley, CA 94720

Contributed by Alexandre J. Chorin, June 19, 2015 (sent for review March 28, 2015; reviewed by Xiantao Li and Juan Restrepo)

Many physical systems are described by nonlinear differential equations that are too complicated to solve in full. A natural way to proceed is to divide the variables into those that are of direct interest and those that are not, formulate solvable approximate equations for the variables of greater interest, and use data and statistical methods to account for the impact of the other variables. In the present paper we consider time-dependent problems and introduce a fully discrete solution method, which simplifies both the analysis of the data and the numerical algorithms. The resulting time series are identified by a NARMAX (nonlinear autoregression moving average with exogenous input) representation familiar from engineering practice. The connections with the Mori–Zwanzig formalism of statistical physics are discussed, as well as an application to the Lorenz 96 system.

discrete approximation | stochastic parametrization | dimension reduction | chaotic systems | NARMAX

There are many time-dependent problems in science, where the equations of motion are too complex for full solution, either because the equations are not certain or because the computational cost is too high, but one is interested only in the dynamics of a subset of the variables. Such problems appear, for example, in weather and climate modeling, e.g., refs. 1, 2; in economics, e.g., ref. 3; in statistical mechanics, e.g., refs. 4, 5; and in the mechanics of turbulent flow, e.g., refs. 6, 7. In these settings it is natural to look for simpler equations that involve only the variables of interest, and then use data to assess the effect of the remaining variables on the variables of interest in some approximate way. In the present paper we focus on stochastic methods for doing this, with data coming either from experimental observations or from prior numerical calculations.

Consider a set of differential equations of the form

$$\frac{d}{dt}x = R(x,y), \quad \frac{d}{dt}y = S(x,y), \qquad [1]$$

where t is the time, $x=(x_1,x_2,\ldots,x_m)$ is the vector of resolved variables, and $y=(y_1,y_2,\ldots,y_\ell)$ is the vector of unresolved variables, where ℓ is not necessarily finite, with initial data $x(0)=\alpha$, $y(0)=\beta$. Consider a situation where this system is too complicated to solve, but where data are available, usually as sequences of measured values of x, assumed here to be observed with negligible observation errors. One can write $R(x,y)$ in the form

$$R(x,y) = R_0(x) + z(x,y), \qquad [2]$$

where R_0 approximates $R(x,y)$ in some sense and is such that one is able to solve the equation

$$\frac{d}{dt}x = R_0(x). \qquad [3]$$

The remainder $z(x,y)=R(x,y)-R_0(x)$ is the contribution of the unresolved variables to the equation for x and must be taken into account. It has been variously called "unresolved tendency" (8, 9), "model error" (10, 11), or "model noise" (12); we call it simply the "noise." In the present paper we do not discuss general methods for constructing R_0; they depend on practical considerations which differ from case to case.

A usual approach to the problem of computing x is to identify z from data (8, 13, 14), i.e., find a concise approximate representation $\hat{z}$ of z that can be evaluated on the computer, and then solve the equation

$$\frac{d}{dt}x = R_0(x) + \hat{z}. \qquad [4]$$

When $\hat{z}$ is a stochastic process, this is a "stochastic parametrization." Eq. **4** is an instance of a dimension-reduced equation, in the sense that it has fewer variables than Eq. **1**. However, this approach has some difficulties. In general the available data are measurements of x, not of z; to find z so that it can be identified one has to use Eq. **2**, and in particular differentiate x numerically, which is generally impractical or inaccurate because z may have high-frequency components or fail to be sufficiently smooth, and because the data may not be available at sufficiently small time intervals (an illuminating analysis in a special case can be found in refs. 15, 16). If one can successfully identify z, Eq. **4** generally becomes a nonlinear stochastic differential system, where in general $\hat{z}$ at a given time depends on earlier values of x and $\hat{z}$ (see the next section), which may be hard to solve with sufficient accuracy (17, 18).

Here we take a different approach. We note that Eqs. **3** and **4** are always solved on the computer, i.e., in discrete form, that the data are always given at a discrete collection of points, and that one wishes to determine x but in general one is not interested in determining z. We can therefore avoid the difficult detour through a continuous z followed by a discretization, by working entirely in a discrete setting as follows. We can pick once and for all a particular accurate discretization of Eq. **3** with a particular time step δ,

$$x^{n+1} = x^n + \delta R_\delta(x^n),$$

where R_δ is obtained, for example, from a Runge–Kutta method, and where n indexes the result after n steps. As in the continuous case, this equation has to be corrected to account for the impact of

Significance

Many problems in science are too large and/or too complex to be fully analyzed by standard methods of computation. We present an approach where such problems are treated statistically, and the statistical analysis and the equations actually solved are discrete rather than continuous. We connect our approach to well-known results in statistical physics, and demonstrate its effectiveness in a widely used test problem.

Author contributions: A.J.C. and F.L. designed research, performed research, contributed new analytic tools, analyzed data, and wrote the paper.

Reviewers: X.L., Pennsylvania State University; and J.R., Oregon State University.

The authors declare no conflict of interest.

[1]To whom correspondence should be addressed. Email: chorin@math.berkeley.edu.

www.pnas.org/cgi/doi/10.1073/pnas.1512080112

the unresolved variables, and here also for the possible inaccuracy of the numerical scheme. We use the data to identify the discrepancy sequence, $z_\delta^{n+1} = (x^{n+1} - x^n)/\delta - R_\delta(x^n)$, which is available from x data without approximation. This is equivalent to identifying the following reduced system:

$$x^{n+1} = x^n + \delta R_\delta(x^n) + \delta z_\delta^{n+1}, \qquad [\mathbf{5}]$$

which constitutes a discrete stochastic parametrization.

We assume, as one does in the continuous case, that the system under consideration is ergodic, so that its long-time statistics are stationary. The sequence z_δ^n becomes a stationary time series, which we represent by the widely used NARMAX (nonlinear autoregression moving average with exogenous inputs) representation, with x as an exogenous input. This representation makes it possible to integrate the numerical scheme into the reduced equations, and to take into account efficiently the non-Markovian features of the reduced system as well as model and numerical errors. There is no stochastic differential system to solve. It is important to note that identifying z_δ can be very different from identifying the continuous z. The question, in what sense does z_δ approximate z, is not relevant, because the goal is to calculate x, and z_δ is sufficient for the purpose. Note that z_δ should be a good approximation of z if Eq. **3** can be effectively approximated by a first-order Euler scheme. For practical purposes, the discrete stochastic parametrization accomplishes everything that would be accomplished by a successful continuous parametrization followed by an accurate approximation.

NARMAX Representation

We represent the discrete-time process z_δ in the reduced system [**5**] as a time series in the form of a variant of the NARMAX representation (19–22). The generality of the NARMAX approach will be increasingly important as model reduction methods are applied to more complex problems. To simplify notation, from now on we drop the subscript δ in z_δ that distinguishes it from the remainder z.

Eq. **5** becomes

$$x^{n+1} = x^n + \delta R_\delta(x^n) + \delta z^{n+1}, \quad z^{n+1} = \Phi^{n+1} + \xi^{n+1}, \qquad [\mathbf{6}]$$

for $n = 1,2, \ldots$, where the ξ^{n+1} are independent Gaussian random variables with mean zero and variance σ^2, and Φ^n is the sum:

$$\Phi^n = \mu + \sum_{j=1}^{p} a_j z^{n-j} + \sum_{j=1}^{r}\sum_{i=1}^{s} b_{i,j} Q_i\left(x^{n-j}\right) + \sum_{j=1}^{q} c_j \xi^{n-j}. \qquad [\mathbf{7}]$$

μ, σ^2 and $\{a_j, b_{i,j}, c_j\}$ are parameters to be inferred from data, and the exogenous inputs Q_i, $i = 1, \ldots, s$ are functions to be determined; to simplify the notations, Eq. **7** has been written as if Eq. **6** were scalar. This is the NARMAX representation of x and z. In Eq. **7**, the terms in z are the autoregression part of order p, the terms in ξ are the moving average part of order q. Note that if we substitute $z^n = (x^n - x^{n-1})/\delta - R_\delta(x^{n-1})$ into [**7**] and the second equation in [**6**], we obtain a NARMAX representation for x. A suitable choice of the functions Q_i will be discussed in the context of a specific example and will connect the representation to the approximation of Eq. **3**.

To implement the NARMAX representation, one has to determine its structure and estimate the parameters. Although NARMAX has been widely studied (see, e.g., refs. 19, 23–26 and references therein), one should use the earlier work with caution, especially in the detection of structure by least-squares–based methods, because in the standard NARMAX, unlike here, the exogenous process is independent of the noise process. Suppose one has selected a structure, that is, chosen the functions Q_i and the orders (p,r,s,q) in the representation [**7**]. Because the representation is linear in the parameters $\theta = (\mu, a_j, b_{i,j}, c_j, \sigma^2)$, these parameters can be estimated using conditional likelihoods as follows. The sequence $\{z^n\}$ for $n = 1,2, \ldots, N$ can be computed from the data using the first equation in [**6**]. Once the values of $\xi^1, \ldots, \xi^q$ are known, the noise sequence ξ^n for $q+1 \le n \le N$ can be computed from $\xi^n = z^n - \Phi^n$. This leads to the conditional log-likelihood of the observations x^n for $q+1 \le n \le N$ (up to a constant):

$$l(\theta|\xi^1, \ldots, \xi^q) = -\sum_{n=q+1}^{N} \frac{\left|z^n - \Phi^n\right|^2}{2\sigma^2} - \frac{N-q}{2}\ln\sigma^2.$$

The log-likelihood $l(\theta|\xi^1, \ldots, \xi^q)$ is quadratic in the parameters other than σ^2, its gradient can be easily computed, and the maximum likelihood estimator (MLE) can be obtained by standard gradient-based optimization methods, such as the quasi-Newton method. If the reduced system is ergodic, the MLE is asymptotically consistent, and the initial values of $\xi^1, \ldots, \xi^q$ do not affect the result (22, 25). For convenience, we set $\xi^1 = \cdots = \xi^q = E[\xi^1] = 0$.

We remark that for the above Gaussian likelihood, the MLE is the same as the least-squares estimator for the parameters, which has been widely used in control (23, 24). As shown in ref. 23, the above estimation procedure can be made recursive; also, the noise sequence need not be Gaussian.

The detection of the representation's structure, however, is less straightforward, as is well-known (19, 21). Because in our problem the exogenous processes are not independent of the noise, popular techniques such as orthogonal least-squares and error reduction ratios (see, e.g., ref. 19 and references therein), do not work. We use the following criteria for selecting the structure of the representation: (*i*) it should fit the long-term statistics of the resolved variables, such as the mean and autocorrelation function; (*ii*) as the size of the data increases, the estimated parameters should converge; and (*iii*) the estimated parameters should reflect features of the resolved variables, such as symmetry properties.

It is of interest to relate the NARMAX representation to the Mori–Zwanzig (MZ) formalism (4, 5, 27, 28). The overall setting in the MZ representation is the same as here: one has data α for the x variables, and one samples data β for y from a given initial probability density function (pdf). The MZ formalism creates the approximation $R_0(x)$ in Eq. **3** by conditional expectation:

$$R_0(x) = E[R(x,y)|x],$$

where $E[a|b]$ is the expected value of a with respect to the initial measure given b. When the system is ergodic and the initial pdf for β is the invariant measure conditioned by α, this is the best least-squares approximation of $R(x,y)$ by a function of x. The MZ formalism then yields an expression for $z(x,y)$ in Eq. **2** as a sum of a noise and a non-Markovian memory/dissipation term, corresponding to ξ^{n+1} and Φ^{n+1} in Eq. **7**; note that in [**3**] R_0 is not restricted to the MZ recipe. The MZ expressions are exact, and prove the need for the representation of z to take the memory into account by including information from earlier steps.

Once the initial data $y(0) = \beta$ have been sampled, the MZ equations are deterministic; the MZ formalism proposes to follow in time one particular sample path of the system. For a chaotic system, such as the one discussed in the next section, this may not be computationally feasible. The representation here looks for sample paths of a stationary stochastic process whose statistics agree with the statistics defined by the equations of motion. This is a related but less ambitious and more feasible task.

The evaluation of the various terms in the MZ formalism is difficult; as far as we know, there is only one case in the literature (29) where it has been successfully carried out in full for a nonlinear problem that is not completely trivial. The MZ formalism is a useful starting point for analytic approximations (29, 30), but it is difficult to use it to construct reduced models from data when suitable analytic approximations are not available. The formalism here is a more tractable way to use data for dimension reduction, and generalizes the MZ formalism to a broader class of approximations.

There is a large literature on data-based dimension reduction. In refs. 8, 31, the noise z is represented as the sum of an approximating polynomial in x obtained by regression and a one-step autoregression; the details are in the next section where we compare our results to those in refs. 8, 31. The shortcomings of this representation as a general tool are that it does not allow z to remember past values of x, and that the autoregression term is not necessarily small, making it difficult to solve the equations accurately. Furthermore, in refs. 8, 31, numerical values of a continuous z are obtained by finite differences. In refs. 9, 14, the noise is represented by a conditional Markov chain that depends on both current and past values of x; the Markov chain is deduced from data by binning and counting, assuming that exact observations of z are available. It should be noted that the Markov chain representation is intrinsically discrete, making this work close to ours in spirit. In ref. 32 the noise is treated as continuous and represented in a form that is partly analogous to the NARMAX representation once one translates from the continuum to the grid. An earlier construction of a reduced approximation can be found in ref. 12, where the approach was not yet fully discrete. Other interesting related work can be found in refs. 33–36.

Lorenz 96 Equations

The Lorenz 96 equations (37) are a set of chaotic differential equations that is often used as a metaphor for the atmosphere. It has been widely used as a test bench for various dimension reduction and stochastic parametrization methods (8, 9, 14, 31, 38). Following ref. 14, we use the following formulation of the equations:

$$\frac{d}{dt}x_k = x_{k-1}\left(x_{k+1} - x_{k-2}\right) - x_k + F + z_k,$$

$$\frac{d}{dt}y_{j,k} = \frac{1}{\varepsilon}\left[y_{j+1,k}\left(y_{j-1,k} - y_{j+2,k}\right) - y_{j,k} + h_y x_k\right],$$

with $z_k = h_x / J \sum_j y_{j,k}$, and $k = 1, \ldots, K$, $j = 1, \ldots, J$. The indices are cyclic, $x_k = x_{k+K}, y_{j,k} = y_{j,k+K}$ and $y_{j+J,k} = y_{j,k+1}$. The system is invariant under spatial translations, and the statistical properties are identical for all x_k. The formulation here is equivalent to the original formulation by Lorenz (e.g., refs. 9, 38); the parameter ε measures the time-scale separation between the resolved variables x_k and the unresolved variables $y_{j,k}$. We set $\varepsilon = 0.5$, so that there is no significant time-scale separation between resolved and unresolved processes, as is both more realistic and more difficult to handle for parametrizations (see ref. 38 and references therein). The other parameters are $K = 18$, $J = 20$, $F = 10$, $h_x = -1$, and $h_y = 1$. The ergodicity of the Lorenz 96 system has been numerically verified in earlier work (38) and we do not revisit this issue.

In the following section we present numerical results produced by our NARMAX scheme and compare them to those in ref. 8 labeled POLYAR (polynomial regression and autoregression). We do not compare with the results of ref. 14 because they require exact observations of z. In ref. 8, the continuous z is estimated by finite differences:

$$z_k(t) \approx \frac{x_k(t+\delta) - x_k(t)}{\delta} - x_{k-1}\left(x_{k+1} - x_{k-2}\right) + x_k - F.$$

Then a polynomial regression of the form $z_k(t) = P(x_k(t)) + \eta_k(t)$ is used to fit the data $\{x_k(n\delta), z_k(n\delta)\}$, where $P(x)$ is an approximating polynomial obtained by least squares, and $\eta_k(t)$ is a one-step autoregression [AR(1)] with parameters estimated from the time series $z_k(n\delta) - P(x_k(n\delta))$, for $n = 1,2, \ldots$. This leads to the following reduced stochastic equation:

$$\frac{d}{dt}x_k = x_{k-1}\left(x_{k+1} - x_{k-2}\right) - x_k + F + P(x_k) + \eta_k, \qquad \textbf{[8]}$$

where η_k is an autoregression of the form

$$\eta_k(t+\delta) = \phi\eta_k(t) + \sigma\xi_k(t), \qquad \textbf{[9]}$$

where ϕ, σ are constants deduced from the data, and the $\xi_k(t)$ are independent identically distributed Gaussian random variable with mean zero and variance 1, for each component $k = 1, \ldots, K$ of the equation. This reduced system is solved as follows: given the current time vectors $(\eta_k(t), x_k(t))$, the next time step $\eta_k(t+\delta)$ is calculated from [**9**], and then $x_k(t+\delta)$ is computed by integrating [**8**] by a fourth-order Runge–Kutta method, with $\eta(t)$ kept constant during each time step.

In the NARMAX scheme, we use the representation [**7**], choosing one of the functions $Q_i(x)$ to be $R_\delta(x)$ from the approximation [**3**] and the others to be powers of x. This connects the numerical scheme with the representation of the noise. We select the structure and estimate the parameters as described earlier. The parameters are the same for each spatial component, reflecting the spatial symmetry of the equation. Each component of z remembers only its own past and the past of the corresponding component of x. The term Φ^n in Eq. **7** becomes

$$\begin{aligned}\Phi^n = \mu + \sum_{j=1}^{p} a_j z^{n-j} + \sum_{j=1}^{r}\sum_{l=1}^{d_x} b_{j,l}\left(x^{n-j}\right)^l \\ + \sum_{j=1}^{s}\sum_{l=1}^{d_R} c_{j,l}\left(R_\delta\left(x^{n-j}\right)\right)^l + \sum_{j=1}^{q} d_j \xi^{n-j}.\end{aligned} \qquad \textbf{[10]}$$

The determination of the numerical parameters in this representation is part of the calculation and the values used are listed in the next section.

Numerical Results

In the numerical runs, we generate data for x_k by integrating the full Lorenz 96 system with parameters $(\varepsilon, K, J, F, h_x, h_y) =$

Table 1. Estimated parameters in the POLYAR system

δ	5*	4	3	2	1	0	ϕ	σ
0.01	−0.00002	0.0004	−0.0002	−0.0258	−0.3567	0.0529	0.9948	0.9397
0.05	−0.00003	0.0009	−0.0035	−0.0137	−1.0030	1.3919	0.7626	11.3857

*The column labeled j for $j = 0, \ldots, 5$ contains the coefficient of x to the power j.

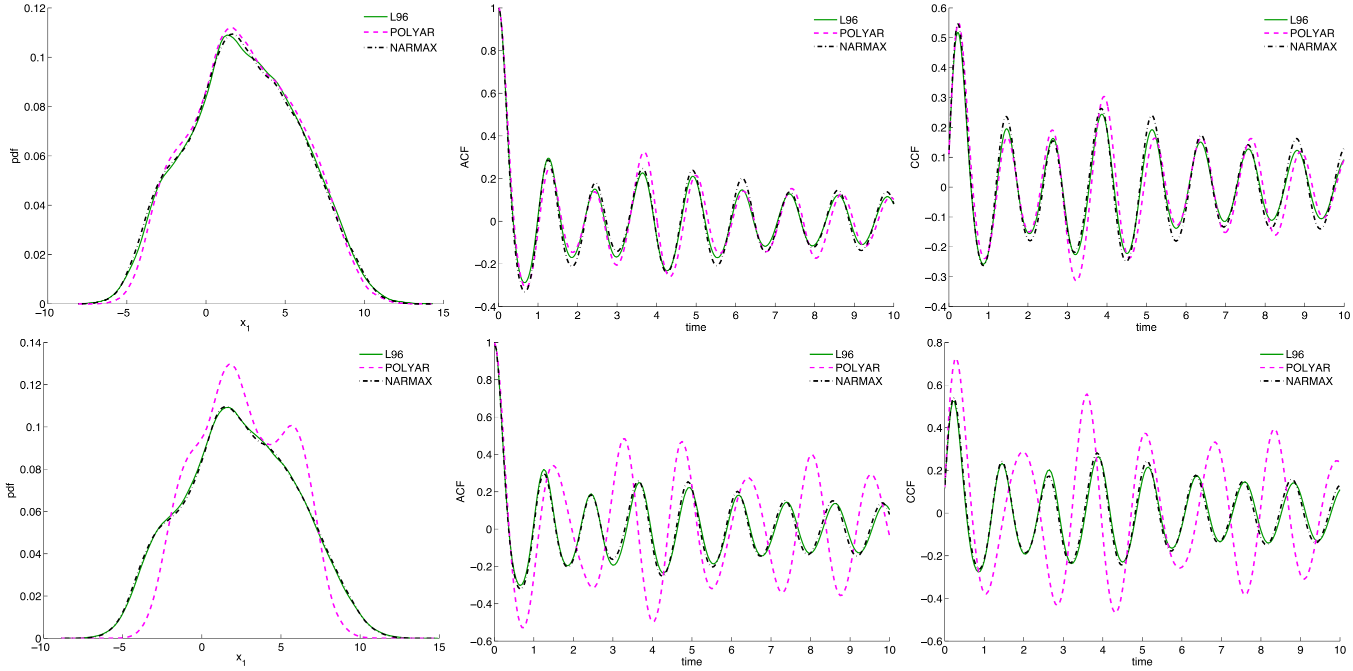

Fig. 1. pdf, ACFs, and CCFs of x_1, produced by the full Lorenz 96 model, the POLYAR system, and the NARMAX system. (*Top*) Case $\delta = 0.01$; (*Bottom*) case $\delta = 0.05$. These statistics are better reproduced by the NARMAX system than the POLYAR system.

$(0.5, 18, 20, 10, -1, 1)$, using a fourth-order Runge–Kutta method with time step 0.001. We consider two cases: one in which the observations are made at intervals of $\delta = 0.01$ and one in which they are made at intervals $\delta = 0.05$; the first case corresponds to a case discussed in refs. 9, 14; in the second case, the data are slightly sparser. In each case, we make observations at $N = 5 \times 10^5$ points; this requires that the full system be integrated for 5,000 and 25,000 time units, respectively. In each case, R_δ comes from the fourth-order Runge–Kutta method.

For the POLYAR equations of [**8**], a fifth-order polynomial is used for both observation settings. Increasing the order further produces small coefficients for the higher degree terms, which do not reduce the variance of noise. The estimated parameters, i.e., the coefficients of the polynomial and the parameters for the autoregression, for the first component x_1 of x are shown in Table 1.

For the NARMAX equation [**10**], we estimated $(p, r, s, q) = (1, 2, 0, 1)$ and $(d_x, d_R) = (1, 0)$ for the case $\delta = 0.01$, and for the case $\delta = 0.05$, $(p, r, s, q) = (1, 1, 1, 0)$ and $(d_x, d_R) = (3, 1)$. The estimated parameters for x_1 are shown in Table 2.

First, we compare the statistics of the solutions generated by the two reduced systems with the statistics of the full system. We integrate the reduced equations in both cases and obtain values at 5×10^5 points. We calculate the following quantities from the reduced equations as well as from the full Lorenz 96 equations: the mean, the SD, the pdf, the Kolmogorov–Smirnov statistic, the autocorrelation function (ACF) of x_1, and the cross-correlation function (CCF) between x_1 and x_2. The pdf of x_1 for the full Lorenz 96 system is well reproduced by both reduced systems when $\delta = 0.01$; Fig. 1 (*Top Left*). In the sparser data case $\delta = 0.05$ the NARMAX equations produce a much better pdf than the POLYAR equations; Fig. 1 (*Bottom Left*). Table 3 displays the mean, the SD, and the Kolmogorov–Smirnov statistic that compare the cumulative distributions of the full Lorenz 96 system with that of the reduced equations. The NARMAX system is more accurate than the POLYAR system in both cases. The ACF and CCF are well reproduced by both reduced systems when $\delta = 0.01$; Fig. 1 (*Middle, Top Right*). When $\delta = 0.05$, however, the ACFs and the CCFs reproduced by the POLYAR miss the amplitude of oscillation and exhibit a phase shift from those of the full Lorenz 96 equations while the NARMAX system remains accurate; Fig. 1 (*Middle, Bottom Right*).

We now investigate how well a reduced system predicts the behavior of the full system by calculating mean trajectories of the reduced systems and comparing them with a true trajectory of the full Lorenz 96 system, as follows. First we integrate the full Lorenz 96 system for $10 \times N_0$ time units, and store the results as N_0 short trajectories of 10 time units each. For each short true trajectory, we perform N_{ens} integrations of the reduced systems over 10 time units, starting all ensemble members from the same several-step initial conditions as the corresponding full solution—several initial steps are needed to initialize η in POLYAR and ξ in NARMAX. We do not introduce artificial perturbations into the initial conditions, because the exact initial conditions for x are known, and by initializing η and ξ from data, we preserve the memory of the system so as to generate better ensemble trajectories. We then average the solutions of the reduced equations in each subinterval and compare these averages with the trajectories of the full system by calculating the root-mean-square error (RMSE)

Table 2. Estimated parameters in the NARMAX model

δ	a_1	$b_{1,1}$	$b_{1,2}$	d_1		μ	σ^2
0.01	0.9782	−0.1271	0.1132	0.9997	—	0.0115	0.0004
	a_1	$b_{1,1}$	$b_{1,2}$	$b_{1,3}$	$c_{1,1}$	μ	σ^2
0.05	0.8879	−0.0712	−0.0002	0.0002	−0.0084	0.0556	0.0284

Table 3. Mean, SD, and the Kolmogorov–Smirnov statistic (D)

	$\delta = 0.01$			$\delta = 0.05$		
	Mean	SD	D	Mean	SD	D
Lorenz 96	2.4506	3.5230	—	2.3978	3.5222	—
POLYAR	2.5335	3.3807	0.0183	2.6031	2.8564	0.0747
NARMAX	2.4113	3.5270	0.0055	2.4293	3.5402	0.0049

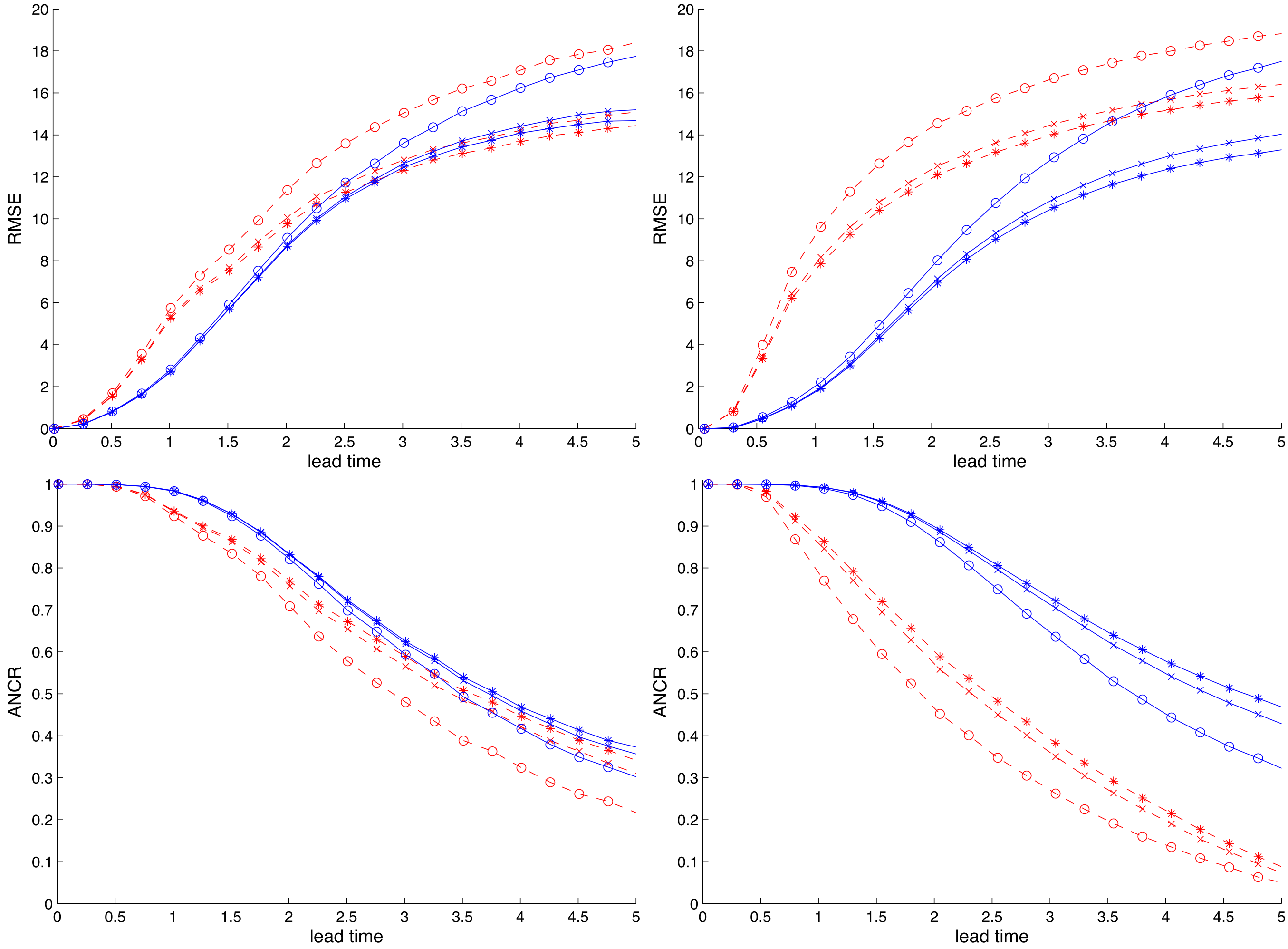

Fig. 2. RMSE and ANCR of ensemble forecasting, produced by the NARMAX system (solid line) and the POLYAR system (dashed line), for different ensemble sizes: $N_{ens} = 1$ (circle marker), $N_{ens} = 5$ (cross marker), and $N_{ens} = 20$ (asterisk marker). (*Left*) Case $\delta = 0.01$; (*Right*) case $\delta = 0.05$. NARMAX has smaller RMSE and larger ANCR than the POLYAR.

and anomaly correlation (ANCR) between them; the former measures the average difference between trajectories whereas the latter measures the average correlation between them; the formulas and their rationale can be found in ref. 14.

Results for RMSE and ANCR are shown in Fig. 2, where we use $N_0 = 10{,}000$, $N_{ens} = 1{,}5{,}20$, and the number of steps where initial conditions are imposed is $n_0 = \max\{1, p, r, s, 2q\} + 1$. In the case $\delta = 0.01$, the NARMAX reduction performs slightly better than the POLYAR reduction. In the case $\delta = 0.05$, the NARMAX reduction provides a significant improvement over the POLYAR reduction. For example, the forecast lead time at which the anomaly correlation drops below 0.6 is extended from $\tau = 2$ to $\tau = 4$ in the case $N_{ens} = 20$.

Conclusions

We have presented a discrete approach to data-based dimension reduction and stochastic parametrization in which the problem is consistently treated as discrete, obviating earlier difficulties in estimating noise from measurements and in approximating reduced continuum equations. Within this discrete approach, we have identified the reduced system via the NARMAX representation. This generalizes earlier work, in particular by making it easier to include memory effects in full. We have tested the resulting algorithm on the Lorenz 96 system of equations, often used as a test bench for dimension reduction schemes; our construction did better than earlier work in reproducing the dynamics and the long-range statistics of the variables of interest, most conspicuously in a problem where the data were sparse. We related our representation to the MZ formalism and suggested that our methods can be used to construct simpler data-based analogs of this formalism. We expect the advantages of our modeling to become even more marked as it is applied to increasingly complex problems.

ACKNOWLEDGMENTS. The authors thank the referees, as well as Prof. Jonathan Goodman, for reading the manuscript carefully and for their helpful suggestions, and Dr. Matthias Morzfeld, Prof. Kevin Lin, Prof. Xuemin Tu, and Prof. Robert Miller for helpful comments and good advice. This work was supported in part by the Director, Office of Science, Computational and Technology Research, US Department of Energy, under Contract DE-AC02-05CH11231, and by the National Science Foundation under Grants DMS-1217065 and DMS-1419044.

1. Kalnay E (2003) *Atmospheric Modeling, Data Assimilation, and Predictability* (Cambridge Univ Press, Cambridge, UK).
2. Wilks DS (2011) *Statistical Methods in the Atmospheric Sciences* (Academic, Oxford, UK).
3. Zeng Y, Wu S, eds (2013) *State Space Models, with Applications in Economics and Finance* (Springer, New York).
4. Evans D, Morris G (1990) *Statistical Mechanics of Nonequilibrium Liquids* (Academic, London).

5. Chorin AJ, Hald OH (2013) *Stochastic Tools in Mathematics and Science* (Springer, New York), 3rd Ed.
6. Bernard P, Wallace J (2002) *Turbulent Flow: Analysis, Measurement, and Prediction* (Wiley, Hoboken, NJ).
7. Chorin AJ (1994) *Vorticity and Turbulence* (Springer, New York).
8. Wilks DS (2005) Effects of stochastic parameterizations in the Lorenz '96 system. *Q J R Meteorol Soc* 131(606):389–407.
9. Kwasniok F (2012) Data-based stochastic subgrid-scale parametrization: An approach using cluster-weighted modelling. *Philos Trans A Math Phys Eng Sci* 370(1962): 1061–1086.
10. Harlim J (2013) Data assimilation with model error from unresolved scales. arXiv: 1311.3579.
11. Berry T, Harlim J (2014) Linear theory for filtering nonlinear multiscale systems with model error. *Proc Math Phys Eng Sci* 470(2167):20140168.
12. Chorin AJ, Hald OH (2014) Estimating the uncertainty in underresolved nonlinear dynamics. *Math Mech Solids* 19(1):28–38.
13. Pavliotis GA, Stuart AM (2007) Parameter estimation for multiscale diffusions. *J Stat Phys* 127(4):741–781.
14. Crommelin D, Vanden-Eijnden E (2008) Subgrid-scale parameterization with conditional Markov chains. *J Atmos Sci* 65(8):2661–2675.
15. Pokern Y, Stuart AM, Wiberg P (2009) Parameter estimation for partially observed hypoelliptic diffusions. *J R Stat Soc, B* 71(1):49–73.
16. Samson A, Thieullen M (2012) A contrast estimator for completely or partially observed hypoelliptic diffusion. *Stochastic Process Appl* 122(7):2521–2552.
17. Kloeden PE, Platen E (1999) *Numerical Solution of Stochastic Differential Equations* (Springer, Berlin), 3rd Ed.
18. Milstein GN, Tretyakov MV (2004) *Stochastic Numerics for Mathematical Physics* (Springer, New York).
19. Billings SA (2013) *Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatiotemporal Domains* (John Wiley & Sons, New York).
20. Brockwell P, Davis R (2002) *Introduction to Time Series and Forecasting* (Springer, New York).
21. Fan J, Yao Q (2003) *Nonlinear Time Series: Nonparametric and Parametric Methods* (Springer, New York).
22. Hamilton JD (1994) *Time Series Analysis* (Princeton Univ Press, Princeton).
23. Ding F, Chen T (2005) Identification of Hammerstein nonlinear ARMAX systems. *Automatica* 41(9):1479–1489.
24. Chen HF (2010) New approach to recursive identification for ARMAX systems. *IEEE Trans Automat Contr* 55(4):868–879.
25. Hannan EJ (1976) The identification and parameterization of ARMAX and state space forms. *Econometrica* 44(4):713–723.
26. Stoffer DS (1986) Estimation and identification of space-time ARMAX models in the presence of missing data. *J Am Stat Assoc* 81(395):762–772.
27. Zwanzig R (1973) Nonlinear generalized Langevin equations. *J Stat Phys* 9(3):215–220.
28. Zwanzig R (2001) *Nonequilibrium Statistical Mechanics* (Oxford Univ Press, New York).
29. Chorin AJ, Hald OH, Kupferman R (2002) Optimal prediction with memory. *Physica D* 166(3):239–257.
30. Stinis P (2013) Renormalized reduced models for singular PDEs. *Commun Appl Math Comput Phys* 8(1):39–66.
31. Arnold HM, Moroz IM, Palmer TN (2013) Stochastic parametrizations and model uncertainty in the Lorenz '96 system. *Philos Trans A Math Phys Eng Sci* 371(1991): 20110479.
32. Majda AJ, Harlim J (2013) Physics constrained nonlinear regression models for time series. *Nonlinearity* 26(1):201–217.
33. Chekroun MD, Kondrashov D, Ghil M (2011) Predicting stochastic systems by noise sampling, and application to the El Niño-Southern Oscillation. *Proc Natl Acad Sci USA* 108(29):11766–11771.
34. Kondrashov D, Chekroun MD, Ghil M (2015) Data-driven non-Markovian closure models. *Physica D* 297:33–55.
35. Duan J, Nadiga B (2007) Stochastic parameterization for large eddy simulation of geophysical flows. *Proc Am Math Soc* 135(4):1187–1196.
36. Du A, Duan J (2009) A stochastic approach for parameterizing unresolved scales in a system with memory. *J Algorithm Comput Technol* 3(3):393–405.
37. Lorenz EN (1996) Predictability: A problem partly solved. *Proc ECMWF Seminar on Predictability* (European Centre for Medium-Range Weather Forecasts, Shinfield Park, Reading, Berkshire, UK), Vol 1, pp 1–18.
38. Fatkulin I, Vanden-Eijnden E (2004) A computational strategy for multi-scale systems with applications to Lorenz'96 model. *J Comput Phys* 200(2):605–638.