Information Theory, Model Error, and Predictive Skill of Stochastic Models for Complex Nonlinear Systems

Dimitrios Giannakis[a,*], Andrew J. Majda[a], Illia Horenko[b]

[a] *Courant Institute of Mathematical Sciences, New York University, New York, NY 10012, USA*
[b] *Institute of Computational Science, University of Lugano, 6900 Lugano, Switzerland*

Abstract

Many problems in complex dynamical systems involve metastable regimes despite nearly Gaussian statistics with underlying dynamics that is very different from the more familiar flows of molecular dynamics. There is significant theoretical and applied interest in developing systematic coarse-grained descriptions of the dynamics, as well as assessing their skill for both short- and long-range prediction. Clustering algorithms, combined with finite-state models for the regime transitions, are a natural way to build such models objectively from observed data in either the true model or an approximate model. The main theme of this paper is the development of new practical criteria to assess the fidelity and predictive skill of such coarse-grained approximations through empirical information theory in stationary and periodically-forced models. These criteria are tested on instructive idealized stochastic models utilizing K-means clustering in conjunction with running-average smoothing of the training data and initial conditions for forecasts. A perspective on these clustering algorithms is explored here with independent interest, where improvement in the information content of finite-state partitions of phase space is a natural outcome of low-pass filtering through running averages. In applications with time-periodic equilibrium statistics, recently developed finite-element, bounded-variation algorithms for nonstationary autoregressive models are shown to strongly influence the fidelity and predictive skill, and substantially improve these features beyond standard autoregressive models.

Keywords: Information theory, Predictability, Model error, Stochastic models, Clustering algorithms, Autoregressive models

1. Introduction

Since the classical work of Lorenz [1–3] and Epstein [4], predictability within dynamical systems has been the focus of extensive study, involving disciplines as diverse as fluid mechanics [5], dynamical-systems theory [6–8], materials science [9–13], atmosphere-ocean science (AOS) [14–28], molecular dynamics (MD) [29–33], econometrics [34, 35], and time series analysis [36–38]. In these and other applications, the dynamics span multiple spatial and temporal scales, take place in phase spaces of large dimension, and are strongly mixing. Yet, despite the complex underlying dynamics, several phenomena of interest are organized around a relatively small number of persistent states (often described in terms of a small set of large-scale observables), which are predictable over timescales significantly longer than suggested by decorrelation times or Lyapunov exponents. Such phenomena often occur in these applications in variables with nearly Gaussian equilibrium statistics [39, 40] and with dynamics that is very different [41] from the more familiar gradient flows, where long-range predictability also often occurs [29, 30]. In other examples, seasonal effects play an important role [23, 34, 42], resulting in time-periodic regime transitions. In either case, revealing predictability in these systems is important both from a practical and a theoretical standpoint. Another issue of key importance is to quantify the fidelity of predictions made with imperfect models when (as is usually the case) the true dynamics of nature cannot be feasibly integrated, or are simply not known [17, 43].

The fundamental perspective adopted here is that predictions in dynamical systems correspond to transfer of information; specifically transfer of information between the initial conditions (which in general are not known completely) and the state of the system at some future time. This opens up the possibility of using the mathematical framework of information theory to characterize both dynamical prediction skill and model error [16–19, 43–50]. The contribution of our work is to further develop and apply this body of knowledge in two important types of predictability problems, which are relevant in many of the disciplinary examples outlined above—namely (i) long-range coarse-grained forecasts in multiscale stochastic dynamical systems; (ii) short- and medium-range forecasts in dynamical systems with time-periodic external forcing.

A major theme prevailing our analysis is to develop techniques and intuition through comparisons of so-called

*Corresponding author. Email: dimitris@cims.nyu.edu

"perfect" or "true" models (which play the role of the inaccessible dynamical system governing the process of interest) with approximate models reflecting our incomplete and/or biased descriptions of the process under study. In (i) the true model will be a three-mode prototype stochastic model featuring physically-motivated dyad interactions [51], and the approximate model will be a nonlinear stochastic scalar model derived via the mode-elimination procedure of Majda et al. [52] (hereafter, MTV). The latter nonlinear scalar model, augmented by suitable time-periodic forcing, will play the role of the true model in (ii), and will be approximated by stationary and non-stationary autoregressive models with external factors (hereafter, ARX models) [23].

The principal results of our study, to our knowledge novel and unanticipated, are that (i) the long-range predictive skill in complex dynamical systems can be revealed through a suitable coarse-grained partition (via data clustering) of the set of initial data, even when short training time series are used; (ii) long-range predictive skill with imperfect models depends simultaneously on the fidelity of these models at asymptotic times, their fidelity during dynamical relaxation to equilibrium, and the discrepancy from equilibrium of forecast probabilities at finite lead times; (iii) nonstationary ARX models can significantly outperform their stationary counterparts in the fidelity of short- and medium-range predictions in challenging nonlinear systems featuring multiplicative noise; (iv) optimal models in the sense of selection criteria based on model complexity [53–55] are not necessarily the models with the highest predictive fidelity. More generally, we demonstrate that information theory provides an objective and unified framework to address these issues. The techniques developed here have potential applications across several disciplines.

In Sec. 2 we briefly review relevant concepts from information theory, and then lay out the general framework for measuring dynamical prediction skill and model error. This framework is applied in Sec. 3 to study long-range coarse-grained forecasts in a time-stationary setting, and in Sec. 4 to study short- and medium-range forecasts in models with time-periodic external forcing. We present our conclusions in Sec. 5.

2. Information theory, predictive skill, and model error

2.1. Predictive skill in a perfect-model environment

We consider the general setting of a stochastic dynamical system

$$d\vec{z} = F(\vec{z},t)\,dt + G(\vec{z},t)\,dW \quad \text{with} \quad \vec{z}\in\mathbb{R}^N, \qquad (1)$$

which is observed through (typically incomplete) observations

$$x(t) = H(\vec{z}(t)), \quad x(t)\in\mathbb{R}^n, \quad n\le N. \qquad (2)$$

Below, $\vec{z}(t)$ will be given either by the three-mode dyad model in Eq. (26) or the nonlinear scalar model in Eq. (28). In other applications (e.g., when dealing with spatially-extended systems [26, 27]), the dimension N of $\vec{z}(t)$ is large. Nevertheless, a number of the essential nonlinear interactions operating in high-dimensional systems are explicitly incorporated in the low-dimensional models studied here. Moreover, as reflected by the explicit dependence of the deterministic and stochastic coefficients in Eq. (1) on time and the state vector, the dynamics of $\vec{z}(t)$ will in general be non-stationary and forced by non-additive noise.

Let $A(t) = A(\vec{z}(t))$ be a prediction observable, i.e., a function of the state vector that is of interest to be predicted. Broadly speaking, the question of dynamical predictability in the setting of Eqs. (1) and (2) may be posed as follows: If we make a measurement $x_0 = x(0) = H(\vec{z}(0))$ at time $t = 0$, how much information have we gained about $A(t)$ at time $t > 0$ in the future? Here, uncertainty in $A(t)$ arises because of both the incomplete nature of the initial data in Eq. (2) and the stochastic component of the dynamical system in Eq. (1). Thus, it is appropriate to describe $A(t)$ via some time-dependent probability distribution $p(A(t) \mid x_0)$ conditioned on the measurement x_0. As the forecast lead-time grows, $p(A(t) \mid x_0)$ will relax towards the equilibrium measure,

$$p_t^{\rm eq} = \lim_{t\to\infty} p(A(t)\mid x_0) = \int dx_0\, p(A(t),x_0), \qquad (3)$$

at which point x_0 contributes no information about $A(t)$.

Above, we have assumed that $p_t^{\rm eq}$ exists and is equal to the marginal distribution of $A(t)$ over all initial measurements. This condition is satisfied by all of the systems studied here (with $p_t^{\rm eq}$ either time-independent, or time-periodic) and many of the applications mentioned in the Introduction. An additional assumption made here when $p_t^{\rm eq}$ is time-independent is that $\vec{z}$ is ergodic,

$$\frac{1}{s}\sum_{i=0}^{s-1} A(\vec{z}(t-i\,\delta t)) \approx \int d\vec{z}\, p^{\rm eq}(\vec{z})A(\vec{z}) \qquad (4)$$

for a large-enough number of samples s. In general, predictability of $A(t)$ may be thought of as the additional information beyond equilibrium conveyed by knowledge of the initial data [16, 18, 26].

The natural mathematical framework to quantify predictability in this context is information theory [44, 48], and, in particular, the concept of relative entropy. The latter is defined as the functional

$$\mathcal{P}(p',p) = \int dA\, p'(A)\log\frac{p'(A)}{p(A)} \qquad (5)$$

between two probability measures, p' and p, and has the attractive properties that (i) it vanishes if and only if $p' = p$, and is positive if $p' \neq p$; (ii) is invariant under general invertible transformations of A. For our purposes, of key

importance is also the so-called Bayesian-update interpretation of relative entropy. This states that if p' is a posterior distribution on A conditioned on some variable x_0 and p is the corresponding prior distribution [which is the case for $p = p^t_{\rm eq}$ in Eq. (3)], then $\mathcal{P}(p(A \mid x_0), p(A))$ measures the additional information beyond p about A gained by having observed x_0. Thus, a natural information-theoretic measure of predictive skill is

$$\mathcal{D}^{x_0}_t = \mathcal{P}(p(A(t) \mid x_0), p^{\rm eq}_t(A)). \tag{6}$$

As one may explicitly verify, the "super-ensemble" expectation value of $\mathcal{D}^{x_0}_t$ over x_0,

$$\mathcal{D}_t = \int dx_0\, p(x_0) \mathcal{D}^{x_0}_t, \tag{7}$$

is also a relative entropy; here between the joint distribution of the prediction observable and the initial data and the product of their marginal distributions. That is, we have the relations

$$\mathcal{D}_t = \mathcal{P}(p(A(t), x_0), p(A(t))p(x_0)) = I(A(t); x_0), \tag{8}$$

where $I(A(t), x_0)$ is known as the mutual information between $A(t)$ and x_0 [15, 18, 26]. Because relative entropy is unbounded from above, it is useful to convert $\mathcal{D}_t$ into a skill score,

$$\delta_t = 1 - \exp(-2\mathcal{D}_t), \tag{9}$$

which lies in the unit interval. Joe [56] shows that the above definition for δ_t is equivalent to a squared correlation measure, at least in problems involving Gaussian variables.

One of the classical results in information theory is that the mutual information between the source and output of a channel measures the rate of information flow across the channel [44, 48]. The maximum of I over the possible source distributions corresponds to the channel capacity. In this regard, an interesting parallel between prediction in dynamical systems and communication across channels is that the combination of dynamical system and measurement apparatus [represented here by Eqs. (1) and (2)] can be thought of as a communication channel with the initial measurements x_0 as input and the prediction observable $A(t)$ as output.

2.2. Quantifying the error in imperfect models

The analysis in Sec. 2.1 was performed in a perfect-model environment. Frequently, however, instead of the true forecast distributions $p(A(t) \mid x_0)$ one has access to distributions $p^M(A(t) \mid x_0)$ derived from an imperfect model,

$$d\vec{z}(t) = F^M(\vec{z}, t)\, dt + G^M(\vec{z}, t)\, dW \tag{10}$$

Such situations arise, for instance, when one cannot afford to feasibly integrate the full dynamical system in Eq. (1) (e.g., MD simulations of biomolecules dissolved in a large number of water molecules), or the laws governing $\vec{z}(t)$ are simply not known (e.g., condensation mechanisms in atmospheric clouds). In other cases, the objective is to develop reliable reduced models for $\vec{z}(t)$ to be used as components of coupled models (e.g., parameterization schemes in climate models [28, 57]). In this context, objective assessments of the error in the model prediction distributions are of key importance, but frequently not carried out in practice [43].

Relative entropy again emerges here as the natural information-theoretic functional for quantifying model error. Now, the analog between dynamical systems and coding theory is with suboptimal coding schemes. In coding theory the expected penalty in the number of bits needed to encode a string assuming that it is drawn from a probability distribution p', when in reality the source probability distribution is p, is given by $\mathcal{P}(p, p')$ evaluated with base-2 logarithms. Similarly, an objective measure of the ignorance or error in an imperfect dynamical model relative to the true model is given by [17, 43, 45, 47]

$$\mathcal{E}^{x_0}_t = \mathcal{P}(p(A(t) \mid x_0), p^M(A(t) \mid x_0)). \tag{11}$$

The above may be aggregated into a super-ensemble measure of model error,

$$\mathcal{E}_t = \int dx_0\, p(x_0) \mathcal{E}^{x_0}_t, \tag{12}$$

with corresponding error score

$$\varepsilon_t = 1 - \exp(-2\mathcal{E}_t), \quad \varepsilon_t \in [0, 1). \tag{13}$$

Consider now a class of imperfect models, $\mathcal{M} = \{M_1, M_2, \ldots\}$ with the corresponding model errors $\mathcal{E}^{\mathcal{M}}_t = \{\mathcal{E}^1_t, \mathcal{E}^2_t, \ldots\}$. An objective criterion for selecting the least-biased model in $\mathcal{M}$ at lead time t is to chose the model with the smallest error in $\mathcal{E}^*_t$ [43]; a choice which will generally depend on t. Alternatively, $\mathcal{E}^{\mathcal{M}}_t$ can be utilized to compute the weights $w_i(t)$ of a mixture distribution $p^*_t = \sum_i w_i(t) p^M_t$ with minimal expected loss of information in the sense of $\mathcal{E}_t$ from Eq. (11) [58]. The latter approach shares certain aspects in common with Bayesian model averaging [59–61], where the weight values w_i are determined by maximum likelihood from the training data. Rather than making multi-model forecasts, in this work our goal is to provide measures to assess the skill and fidelity of a single model given its time-dependent forecast distributions. In particular, one of the key points in the applications of Secs. 3 and 4 ahead is that model assessments should be based on both $\mathcal{E}_t$ and $\mathcal{D}_t$ from Eq. (7).

3. Long-range, coarse-grained forecasts

In our first application, we study long-range coarse-grained predictions in stationary stochastic dynamical systems with metastable low-frequency dynamics. Such dynamical systems arise in applications of wide practical interest (e.g., conformational transitions in MD [29, 30] and climate-regimes in AOS [20, 21, 26, 27, 40]), and are dominated on some coarse-grained scale by switching between

distinct regimes in phase space. Here we demonstrate that long-range predictability may be revealed in these models by employing a suitable partition Ξ of the set of initial data. In this picture, a regime is represented by the integer-valued affiliation S of the initial-data vector x_0 in Eq. (2) to an element of the partition, and is not necessarily related to local maxima in probability density functions (PDFs) [39, 40, 62]. The main tenets here are that (i) S embodies the coarse-grained information relevant to long-range forecasting; (ii) Ξ may be constructed feasibly by data-clustering realizations of the system in equilibrium, thus avoiding the challenging task of ensemble-initialization [63].

More specifically, our strategy is to use the information-theoretic framework of §2 to assess the predictive information associated with Ξ, as well as to quantify the dynamical model error incurred by using imperfect models for the low-frequency dynamics. We develop these techniques in Secs. 3.1 and 3.2, which are followed by an instructive application in Secs. 3.3–3.7 involving nonlinear stochastic models with multiple timescales.

3.1. Coarse-graining phase space to reveal long-range predictability

Our method of phase-space partitioning, described also in Ref. [26], proceeds in two stages: a training stage and prediction stage. The training stage involves taking a dataset

$$\mathcal{X} = \{x((s-1)\,\delta t), x((s-2)\,\delta t), \ldots, x(0)\}, \tag{14}$$

of s samples $x(t)$ from Eq. (2), and computing via data-clustering a collection of K centroids,

$$\Theta = \{\theta_1, \ldots, \theta_K\}, \quad \theta_k \in \mathbb{R}^n. \tag{15}$$

Used in conjunction with a rule for determining S given Θ, the centroids above lead to a mutually-disjoint partition of n-dimensional observation space,

$$\Xi = \{\xi_1, \ldots, \xi_K\}, \quad \xi_k \subset \mathbb{R}^n, \tag{16}$$

such that $S = k$ indicates that the affiliation of the system at time $t = 0$ is with cluster $\xi_k \in \Xi$. In the prediction stage, the cluster-conditional probabilities

$$p_t^k(A) = p(A(t) \mid S = k) \tag{17}$$

for observable $A(t)$ are computed by bin-counting realizations of $A(t)$ and S, which are independent from the dataset in Eq. (14) employed in the training stage. The predictive skill in the true model is then measured via the relative entropy metrics in Eqs. (6) and (7), viz.,

$$\mathcal{D}_t^k = \mathcal{P}(p_t^k, p_{\text{eq}}) \quad \text{and} \quad \mathcal{D}_t = \sum_{k=1}^{K} \pi_k \mathcal{D}_t^k, \tag{18}$$

where $\pi_k = p(S = k)$ is the probability of affiliation with cluster k in equilibrium.

Besides the number of regimes K, our partitioning algorithm has two free parameters. These are temporal windows, Δt and $\Delta\tau$, used to smooth $x(t)$ via running-averaging in the training and prediction stages, respectively. This procedure, which is reminiscent of kernel density estimation methods [64], leads to a two-parameter family of partitions as follows:

First, set an integer $q' \geq 1$, and replace $x(t)$ in Eq. (14) with the averages over a time window $\Delta t = (q'-1)\,\delta t$, i.e.,

$$x^{\Delta t} = \sum_{i=1}^{q'} x(t - (i-1)\,\delta t)/q'. \tag{19}$$

Next, apply K-means clustering [65, 66] to the above coarse-grained training data. This leads to set of coordinates in Eq. (15). In the second, prediction stage, of the procedure, initial data $x(t)$ are collected over an interval $[-\Delta\tau, 0]$ with $\Delta\tau = (q-1)\,\delta t$, and their average $x^{\Delta\tau}$ is computed via an analogous formula to Eq. (19). It is important to note that the initial data in the prediction stage are independent of the the training dataset. The affiliation function S is then given by

$$S = \underset{k}{\text{argmin}}(\|x^{\Delta\tau} - \theta_k^{\Delta t}\|_2); \tag{20}$$

i.e., S depends on both Δt and $\Delta\tau$. By our ergodicity assumption in Eq. (4), the p_t^k from Eq. (17) may be estimated by binning the cluster-conditional samples

$$\mathcal{A}_t^k = \{A(t) : S = k\}. \tag{21}$$

for each $k \in \{1, \ldots, K\}$, given samples of the doublet $\{S, A(t)\}$ over a long-enough time [27]. The key point here is that optimal values for Δt and $\Delta\tau$ (as well as K) maximizing the predictive information content in the partition can be determined *a posteriori* via the relative-entropy measure in Eqs. (18).

3.2. Quantifying the model error in long-range forecasts

Suppose now that instead of the true model one has access to an imperfect model that, as described in Sec. 2.2, produces prediction probabilities

$$p_t^{Mk}(A) = p^M(A(t) \mid S = k), \quad \lim_{t\to\infty} p_t^{Mk} = p_{\text{eq}}^M, \tag{22}$$

which may be systematically biased away from p_t^k in Eq. (22). Here, an obvious candidate measure for predictive skill follows by writing down Eq. (18) with p_t^k replaced by p_t^{Mk}, i.e.,

$$\mathcal{D}_t^{Mk} = \mathcal{P}(p_t^{Mk}, p_{\text{eq}}^M), \tag{23a}$$

$$\mathcal{D}_t^M = \sum_{k=1}^{K} \pi_k^M \mathcal{D}_t^{Mk}, \quad \pi_k^M = p^M(S = k). \tag{23b}$$

The above measures the discrepancy from equilibrium of the prediction probabilities in the model. A major deficiency of the measures in Eqs. (23) is that by being based

solely on PDFs internal to the model they fail to take into account model error (or "ignorance" in the model relative to the truth) [17, 27, 43].

Note, in particular, the distinguished role that the model equilibrium distribution plays in Eq. (23a): If $p^M_{\rm eq}$ differs systematically from the truth, then $\mathcal{D}_t^{Mk}$ conveys false predictive skill at *all* times (including $t = 0$), irrespective of the fidelity of p_t^{Mk} at finite times. This observation leads naturally to the requirement that long-range forecasting models must reproduce the equilibrium statistics of the true model with high fidelity. In the information-theoretic framework of Sec. 2.2, this is expressed as

$$\mathcal{E}_{\rm eq} \ll 1, \quad \text{with} \quad \mathcal{E}_{\rm eq} = \lim_{t\to\infty} \mathcal{E}_t. \tag{24}$$

Here, we refer to the criterion in Eq. (24) as equilibrium consistency; an equivalent condition is called fidelity [67], or climate consistency [27] in AOS work.

Even though equilibrium consistency is a necessary condition for skillful long-range forecasts, it is not a sufficient condition. In particular, the model error at finite lead-time t, expressed from Eq. (12) as

$$\mathcal{E}_t^k = \mathcal{P}(p_t^k, p_t^{Mk}), \tag{25}$$

may be large, despite eventually decaying to a small value at asymptotic times. Thus, long-range forecasting models must simultaneously satisfy $\mathcal{E}_t \ll 1$ at the forecast lead-time of interest, as well as Eq. (24) in equilibrium. If both of these conditions are met, then the $\mathcal{D}_t^M$ metric in Eq. (23b) can be used to measure genuine gain of information relative to the true equilibrium distribution. In summary, our analysis indicates that assessments of long-range predictions with imperfect models should take into consideration all of $\mathcal{E}_{\rm eq}$, $\mathcal{E}_t$, and $\mathcal{D}_t^M$.

3.3. The three-mode dyad model

Here, we consider that the perfect model of Eq. (1) is a three-mode nonlinear stochastic model in the family of prototype models developed by Majda et al. [68], which mimic the structure of non-linear interactions in high-dimensional fluid-dynamical systems. Among the components of the state vector $\vec{z} = (x, y_1, y_2)$, x is intended to represent a slowly-evolving scalar variable accessible to observation, whereas the remaining modes, y_1 and y_2, act as surrogate variables for the unresolved degrees of freedom. The unresolved modes are coupled to x linearly and via a dyad interaction between x and y_1, and x is also driven by external forcing (assumed, for the time being, constant). Specifically, the governing stochastic differential equations are

$$dx = (Ixy_1 + L_1y_1 + L_2y_2 + F + Dx)\, dt \tag{26a}$$
$$dy_1 = \left(-Ix^2 - L_1x - \gamma_1\epsilon^{-1}y_1\right)\, dt + \sigma_1\epsilon^{-1/2}\, dW_1, \tag{26b}$$
$$dy_2 = \left(-L_2x - \gamma_2\epsilon^{-1}y_2\right)\, dt + \sigma_2\epsilon^{-1/2}\, dW_2, \tag{26c}$$

where where $\{W_1, W_2\}$ are independent Wiener processes [69, 70], and the parameters I, $\{D, L_1, L_2\}$, and F respectively measure the the dyad interaction, the linear couplings, and the external forcing. The parameter ϵ controls the time-scale separation of the dynamics of the slow and fast modes, with the fast modes evolving infinitely fast relative to the slow mode in the limit $\epsilon \to 0$. This model, as well as the associated reduced scalar model in Eq. (28) ahead, have been used as prototype models to develop methods based on the fluctuation-dissipation theorem (FDT) for assessing the low-frequency climate response on external perturbations (e.g., CO_2 forcing) [51].

Representing the imperfect model in Eq. (10) is a scalar stochastic model associated with the three-mode model in the limit $\epsilon \to 0$. This reduced version of the model is particularly useful in exposing in a transparent manner the influence of the unresolved modes when there exists a clear separation of timescales in their respective dynamics (i.e., when ϵ is small). As follows by applying the MTV mode-reduction procedure [52, 71] to the coupled system in Eqs. (26), the reduced model is governed by the nonlinear stochastic differential equation

$$dx = (F + Dx)\, dt \tag{27a}$$
$$+ \epsilon\left(\frac{\sigma_2^2 I L_1}{2\gamma_1^2} + \left(\frac{\sigma_1^2 I^2}{2\gamma_1^2} - \left(\frac{L_1^2}{\gamma_1} + \frac{L_2^2}{\gamma_2}\right)\right)x\right.$$
$$\left. - \frac{2IL_1}{\gamma_1}x^2 - \frac{I^2}{\gamma_1}x^3\right)\, dt \tag{27b}$$
$$+ \epsilon^{1/2}\frac{\sigma_1}{\gamma_1}(Ix + L_1)\, dW_1 \tag{27c}$$
$$+ \epsilon^{1/2}\frac{\sigma_2}{\gamma_2}L_2\, dW_2. \tag{27d}$$

The above may also be expressed in the form

$$dx = (\tilde{F} + ax + bx^2 - cx^3)\, dt + (\alpha - \beta x)\, dW_1 + \sigma\, dW_2 \tag{28}$$

with the parameter values

$$\begin{aligned} \tilde{F} &= F + \epsilon\frac{\sigma_1^2 I L_1}{2\gamma_1^2}, \\ a &= D + \epsilon\left(\frac{\sigma_1^2 I^2}{2\gamma_1^2} - \left(\frac{L_1^2}{\gamma_1} + \frac{L_2^2}{\gamma_2}\right)\right), \\ b &= -\epsilon\frac{2IL_1}{\gamma_1}, \quad c = \epsilon\frac{I^2}{\gamma_1}, \\ \alpha &= \epsilon^{1/2}\frac{\sigma_1 L_1}{\gamma_1}, \quad \beta = -\epsilon^{1/2}\frac{\sigma_1 I}{\gamma_1}, \quad \sigma = \epsilon^{1/2}\frac{\sigma_2 L_2}{\gamma_2}. \end{aligned} \tag{29}$$

Among the terms in the right-hand side of Eq. (27) we identify (i) the bare truncation (27a); (ii) a nonlinear deterministic driving (27b) of the climate mode mediated by the linear and dyad interactions with the unresolved modes; (iii) correlated additive-multiplicative (CAM) noise (27c); (iv) additive noise (27d). Moreover, note the parameter interdependence $\beta/\alpha = c/2b = -I/L_1$. This is a manifestation of the fact that in scalar models of the form in Eq. (27), whose origin lies in multivariate models with

multiplicative dyad interactions, a nonzero multiplicative-noise parameter β is accompanied by a nonzero cubic damping c [68].

A useful property of the reduced scalar model is that its equilibrium PDF, $p_{\rm eq}^M(x)$, may be determined analytically by solving the corresponding time-independent Fokker-Planck equation [40]. Specifically, for the governing stochastic differential equation (27) we have the result

$$p_{\rm eq}^M(x) = \frac{N}{((\beta x-\alpha)^2+\sigma^2)^{\tilde{a}}} \exp\left(\tilde{d}\,\mathrm{atan}\left(\frac{\beta x-\alpha}{\sigma}\right)\right) \times \exp\left(\frac{\tilde{b}x-\tilde{c}x^2}{B^4}\right), \quad (30)$$

expressed in terms of the parameters

$$\tilde{a} = 1 - \frac{-3\alpha^2 c + a\beta^2 + 2\alpha b\beta + c\sigma^2}{\beta^4},$$
$$\tilde{b} = 2b\beta^2 - 4c\alpha\beta, \quad \tilde{c} = c\beta^2$$
$$\tilde{d} = \frac{d'}{\sigma} + d''\sigma, \quad d' = \frac{2\alpha^2 b\beta - 2\alpha^3 c + 2\alpha a\beta^2 + 2\beta^3\tilde{F}}{\beta^4},$$
$$d'' = \frac{6c\alpha - 2b\beta}{\beta^4}. \quad (31)$$

Eq. (30) reveals that cubic damping has the important role of suppressing the power-law tails of the PDF arising when CAM noise acts alone, which are not compatible with climate data [39, 40].

3.4. Parameter selection and equilibrium statistics

We adopt the model-parameter values chosen in Ref. [51] in work on the FDT, where the three-mode dyad model and the reduced scalar model were used as test models mimicking the dynamics of large-scale global circulation models. Specifically, we set $I = 1$, $\sigma_1 = 1.2$, $\sigma_2 = 0.8$, $D = -2$, $L_1 = 0.2$, $L_2 = 0.1$, $F = 0$, $\gamma_1 = 0.1$, $\gamma_2 = 0.6$, and ϵ equal to either 0.1 or 1. The corresponding parameters of the reduced scalar model are listed in Table 1. The $\tilde{b}$ and $\tilde{c}$ parameters, which govern the transition from exponential to Gaussian tails of the equilibrium PDF in Eq. (30), have the values $(\tilde{b},\tilde{c}) = (-0.0089, 0.0667)$ and $(\tilde{b},\tilde{c}) = (-0.8889, 6.6667)$ respectively for $\epsilon = 0.1$ and $\epsilon = 1$. For the numerical integrations of the models, we used an RK4 scheme for the deterministic part of the governing equations and a forward-Euler or Milstein scheme for the stochastic part [70], respectively for the three-mode and reduced models. Throughout, we use a timestep equal to 10^{-4} natural time units and an initial equilibration time equal to 2000 natural time units [cf. the O(1) decorrelation times in Table 2].

As shown in Fig. 1, with this choice of parameter values the equilibrium PDFs for x are unimodal and positively skewed in both the three-mode and scalar models. For positive values of x the distributions decay exponentially (the exponential decay persists at least until the 6σ level),

Table 1: Parameters of the scalar stochastic model in Eq. (28) for $\epsilon = 0.1$ and $\epsilon = 1$

ϵ	$\tilde{F}$	a	b	c	α	β	σ
0.1	0.04	−1.809	−0.067	0.167	0.105	−0.634	0.063
1	0.4	−0.092	−0.667	1.667	0.333	−2	0.2

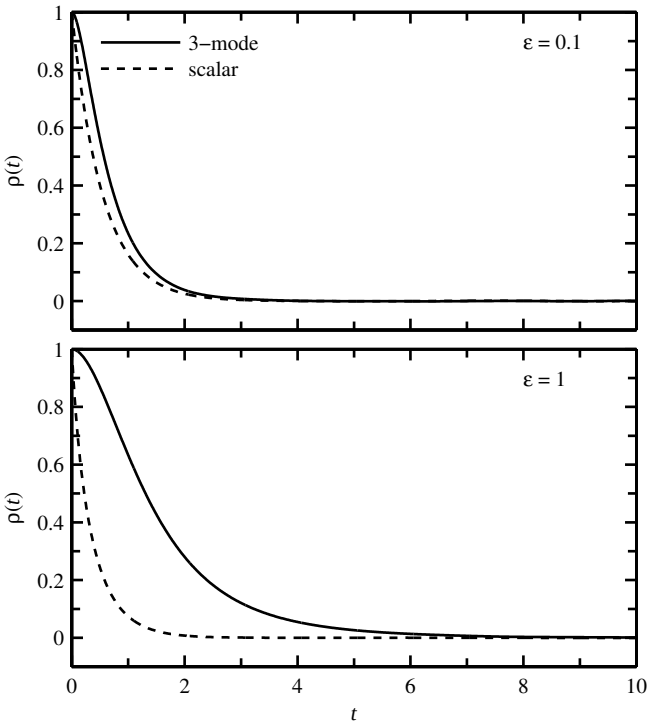

Figure 2: Normalized autocorrelation function, $\rho(t) = \int_0^T dt'\, x(t)x(t'+t)/(T\,\mathrm{var}(x))$, of mode x in the three-mode and reduced scalar models with $\epsilon = 0.1$ and 1. The values of the corresponding correlation time, $\tau_c = \int_0^T dt\,\rho(t)$, are listed in Table 2.

but, as indicated by the positive $\tilde{c}$ parameter in Eq. (30), cubic damping causes the tail distributions to eventually become Gaussian. The positive skewness of the distributions is due to CAM noise with negative β parameter (see Table 1), which tends to amplify excursions of x towards large positive values. In all of the considered cases, the autocorrelation function exhibits a nearly monotonic decay to zero, as shown in Fig. 2.

The marginal equilibrium statistics of the models are summarized in Table 2. According to the information in that table, approximately 99.5% of the total variance of the $\epsilon = 0.1$ three-mode model is carried by the unresolved modes, y_1 and y_2; a typical scenario in AOS applications. Moreover, the equilibrium statistical properties of the reduced model are in good agreement with the three-mode model. As expected, that level of agreement does not hold in the case of the $\epsilon = 1$ models, but, intriguingly, the probability distributions appear to be related by similarity transformations [51].

Table 2: Equilibrium statistics of the three-mode and reduced scalar models for $\epsilon \in \{0.1, 1\}$. Here, the skewness and kurtosis are defined respectively as $\mathrm{skew}(x) = (\langle x^3\rangle - 3\langle x^2\rangle\bar{x} + 2\bar{x}^3)/\,\mathrm{var}(x)^{3/2}$ and $\mathrm{kurt}(x) = (\langle x^4\rangle - 4\langle x^3\rangle\bar{x} + 6\langle x^2\rangle\bar{x}^2 - 3\bar{x}^4)/\,\mathrm{var}(x)^2$; for a Gaussian variable with zero mean and unit variance they take the values $\mathrm{skew}(x) = 0$ and $\mathrm{kurt}(x) = 3/4$. The quantity τ_c is the decorrelation time defined in the caption of Fig. 2.

	$\epsilon = 0.1$		$\epsilon = 1$	
	x (three-mode)	x (scalar)	x (three-mode)	x (scalar)
$\bar{x}$	0.0165	0.0219	0.0461	0.163
$\mathrm{var}(x)$	0.00514	0.00561	0.0278	0.128
$\mathrm{skew}(x)$	1.4	1.38	3.01	2.22
$\mathrm{kurt}(x)$	7.3	7.16	18.2	10.4
τ_c	0.727	0.552	1.65	0.366
	y_1	y_2	y_1	y_2
$\bar{y}_i$	$-4.22\mathrm{E}-05$	0.000355	-0.0671	-0.0141
$\mathrm{var}(y_i)$	1.2	0.801	1.1	0.788
$\mathrm{skew}(y_i)$	-0.000593	-0.000135	-0.0803	0.0011
$\mathrm{kurt}(y_i)$	3	3	2.96	3
τ_c	0.17	0.254	1.41	2.45

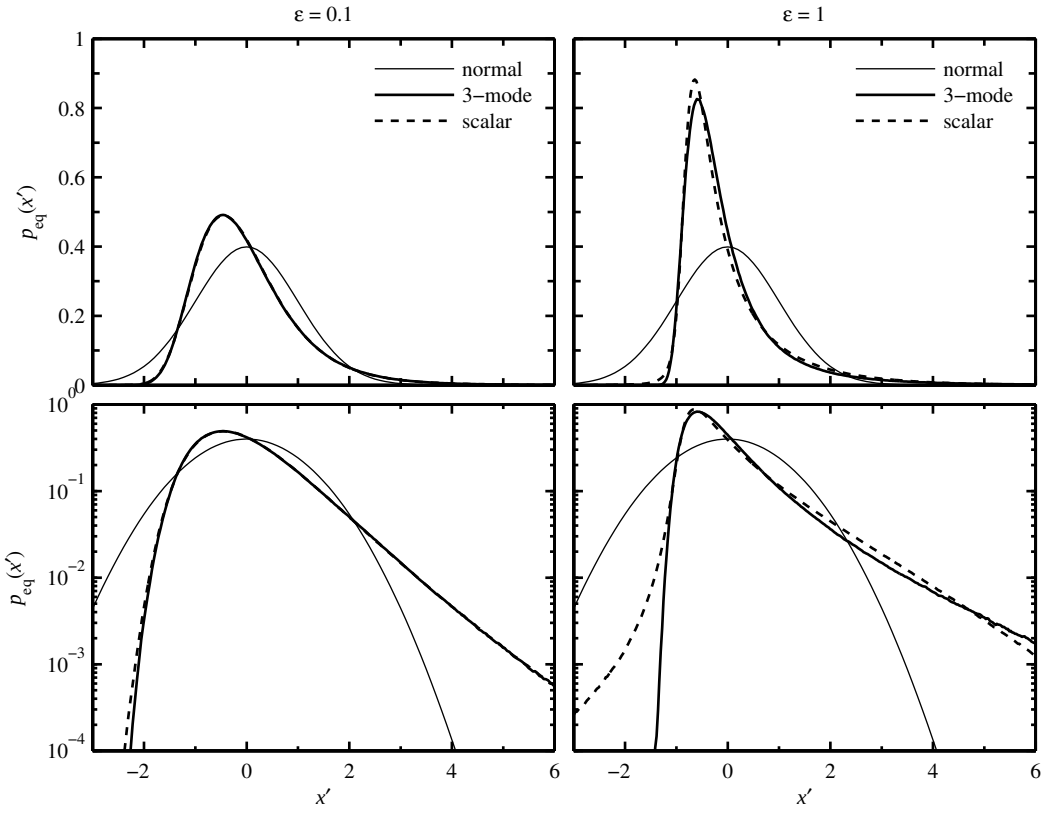

Figure 1: Equilibrium PDFs of the resolved mode x of the three-mode (thick solid lines) and scalar models (dashed lines) for $\epsilon = 0.1$ (left-hand panels) and $\epsilon = 1$ (right-hand panels). Shown here is the marginal PDF of the standardized variable $x' = (x - \bar{x})/\,\mathrm{stdev}(x)$ in linear (top panels) and logarithmic scales (bottom-row panels). The Gaussian distribution with zero mean and unit variance is also plotted for reference in a thin solid line.

3.5. Revealing predictability beyond correlation times

First, we study long-range predictability in a perfect model environment. As remarked earlier, we consider that only mode x of the three-mode system in Eq. (26) is accessible to observations, and therefore carry out the clustering procedure in Sec. 3.1 using that mode alone. For each of the three-mode and scalar models with $\epsilon = 0.1$ and 1, we took a training time series of length $T = 400$, sampled every $\delta t = 0.01$ time units (i.e., $T = s\,\delta t$ with $s = 40{,}000$), and coarse-grained using a running-average interval $\Delta t = 1.6 = 160\,\delta t$. Thus, we have $T \simeq 550\tau_c$ and $\Delta t \simeq 2.2\tau_c$ for $\epsilon = 0.1$; and $T \simeq 250\tau_c$ and $\Delta t \simeq \tau_c$ for $\epsilon = 1$ (see Table 2). In each case, the length of the time series used to estimate the cluster-conditional PDFs in the prediction stage was $T' = 6400$, and the running-average window $\Delta\tau = \delta t = 0.01$; i.e., no coarse-graining is performed in the prediction stage.

In Fig. 3(a,b) we display the dependence of the superensemble skill score δ_t from Eq. (9) for mode x of the three-mode model on the prediction lead-time t, for partitions with $K \in \{2, \ldots, 5\}$. Also shown in those panels are the exponentials $\delta_t^c = -\exp(-2t/\tau_c)$, decaying at a rate twice as fast than the decorrelation time of mode x. Because the δ_t skill score is associated with squared correlations [56], a weaker decay of δ_t compared with δ_t^c signals predictability in mode x beyond its decorrelation time. This is evident in Fig. 3(a,b), especially for $\epsilon = 1$. The fact that decorrelation times are frequently poor indicators of predictability (or lack thereof) has been noted elsewhere in the literature [25, 26].

Turning now to the reduced scalar model, in Fig. 3(c,d) we show the unit-normalized skill score $\delta_t^M = 1 - \exp(-2\mathcal{D}_t^M)$ determined from the relative entropy metric in Eq. (23b). According to the analysis in Sec. 3.2, when the reduced scalar model is used to make predictions of mode x in the three-mode model, δ_t^M may convey false predictive skill. Deferring that discussion to Sec. 3.7, here we use δ_t^M to point out a prominent difference between the model-intrinsic predictability in the scalar model compared to the three-mode model: As manifested by the rate of decay of the δ_t^M in Fig. 3(c,d), which is faster than δ_t^c, the scalar model lacks predictability beyond correlation time. This is because the deterministic driving of mode x in Eq. (26a) by the unresolved modes, y_1 and y_2, is replaced in Eq. (27) governing the scalar model with a forcing that contains a deterministic component [Eq. (27b)], as well as stochastic contributions [Eqs. (27c) and (27d)]. Evidently, some loss of information takes place in the stochastic description of the x–y interaction, which is reflected in the stronger decay of the δ_t^M metric compared with δ_t.

The significant difference in predictability between the three-mode and scalar model despite their similarities in low-frequency variability (as measured, for instance, by the autocorrelation function in Fig. 2), is a clear example that low-frequency variability does not necessarily translate to predictability. The information-theoretic metrics developed here allow one to identify when low-frequency variability is due to noise or deterministic dynamics.

3.6. Length of the training time series

In the idealized case of an infinitely-long training time series, $T \to \infty$, the cluster coordinates Θ are T-independent for ergodic dynamical systems. However, for finite T the computed values of Θ differ between independent realizations of the training time series. As T becomes small (possibly, but not necessarily, comparable to the decorrelation time of the training time series), one would generally expect the information content of the phase-space partition associated with Θ to decrease. An understanding of the relationship between T and model skill is particularly important in practical applications, where one is frequently motivated and/or constrained to work with short training time series.

Here, we study the influence of T on model skill through the superensemble score δ_t in Eq. (9), evaluated for mode x of the three-model model at prediction time $t = 0$. Effectively, this measures the skill of the clusters Θ in classifying observations of x in equilibrium. Even though the behavior of δ_t for $t > 0$ is not necessarily predetermined by δ_0, at a minimum, if δ_0 becomes small as a result of decreasing T, then it is highly likely that δ_t will be correspondingly influenced.

In Fig. 4 we display δ_0 for representative values of T spaced logarithmically in the interval $0.32 \approx 0.4\tau_c$ to $800 \approx 1100\tau_c$ and cluster number K in the range 2–4. Throughout, the running-average intervals in the training and prediction stages are $\Delta t = 160\,\delta t = 1.6 \approx 2.5\tau_c$ and $\Delta\tau = \delta t$ (note that δ_0 is a decreasing function of $\Delta\tau$ for mode x, but may be non-monotonic in other applications; see, e.g., Ref. [26]). Model skill remains fairly independent of the training time series length down to values of T between 2–3 multiples of the correlation time τ_c, at which point δ_0 begins to decrease rapidly with decreasing T.

The results in Fig. 4 demonstrate that informative partitions of phase space can be computed using training data spanning only a few multiples of the correlation time. This does not mean, however, that such small datasets are sufficient to produce a practical predictive model. In particular, making predictions assumes knowledge of the cluster-conditional probabilities $p_t^k(x)$ in Eq. (17), and estimating those probabilities without significant sampling error generally requires longer time series. Here we do not examine this source of model error, and, as stated above, use throughout an independent time series of length $T' = 6400 \gg T$ to compute the cluster-conditional PDFs empirically.

3.7. Dynamical error in the reduced scalar model

In this section, we assess the model error incurred by using the reduced scalar model to approximate mode x in the three-mode model. This error is measured in superensemble forecasts by the relative entropy $\mathcal{E}_t$ in Eq. (12), or, equivalently by the unit-normalized "score" ε_t in Eq. (13).

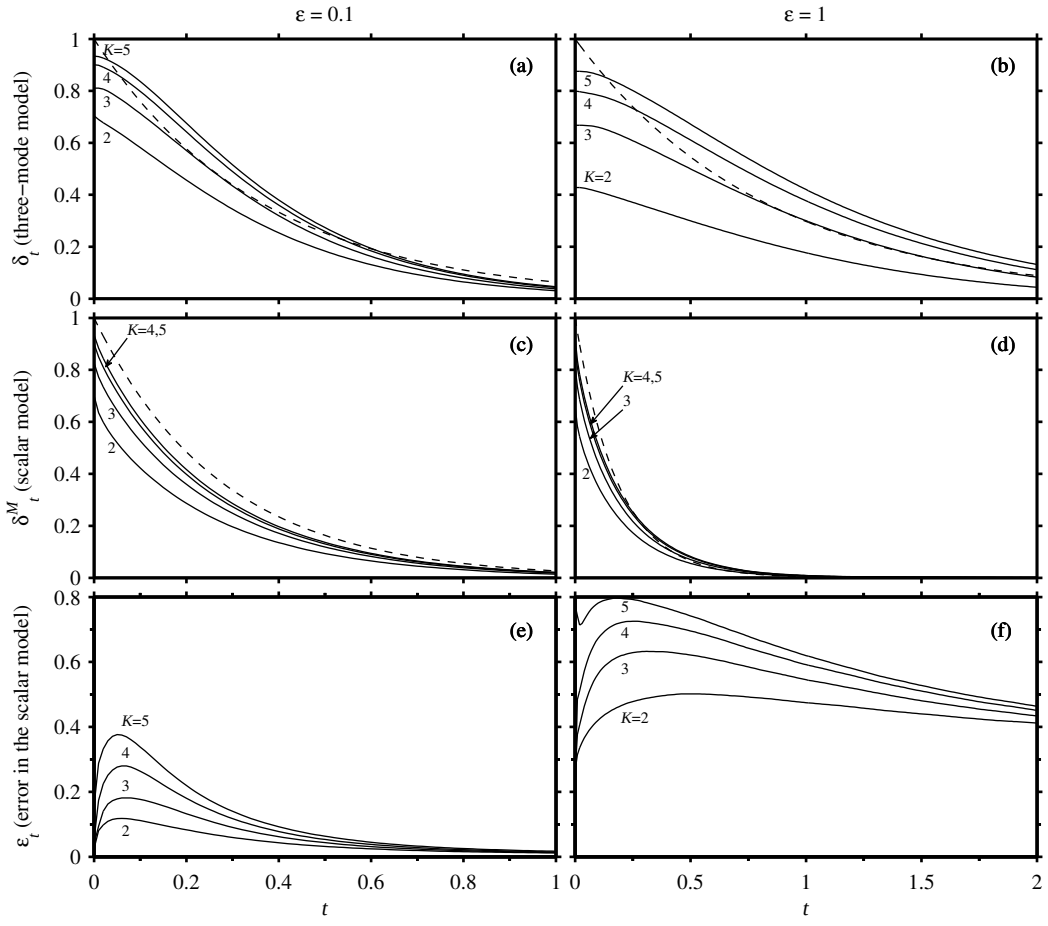

Figure 3: Predictive skill in the three-mode model and model error in the reduced scalar model for phase-space partitions with $K \in \{2, \ldots, 5\}$. Shown here are (a,b) the predictive skill score δ_t for mode x of the three-mode model; (c,d) the discrepancy from equilibrium δ_t^M in the scalar model; (e,f) the normalized error ε_t in the scalar model. The dotted lines in Panels (a–d) are exponential decays $\delta_t^c = \exp(-2t/\tau_c)$ based on half of the correlation time τ_c of mode x in the corresponding model. A weaker decay of δ_t compared to δ_t^c indicates predictability beyond correlation time. Because ε_t in Panel (f) is large at late times, the scalar model with $\epsilon = 1$ fails to meet the equilibrium consistency criterion in Eq. (24). Thus, the δ_t^M score in Panel (d) measures false predictive skill.

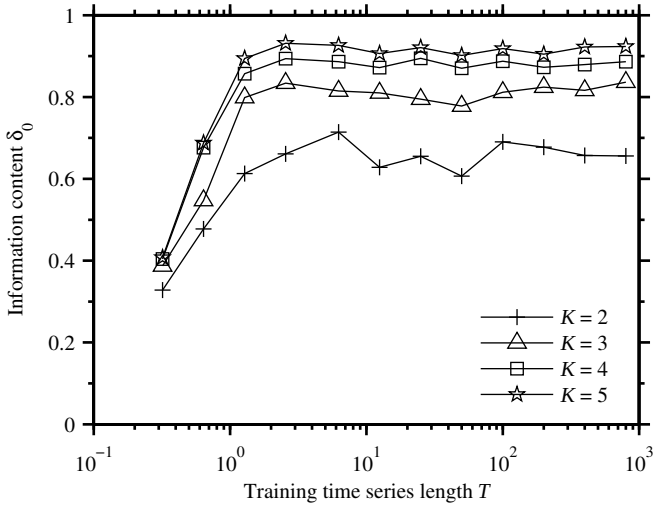

Figure 4: Information content δ_0 in the partitions for mode x of the three-mode dyad model with $\epsilon = 0.1$ as a function of the length T of the training time series. Note the comparatively small gain in information in going from $K = 4$ to 5 clusters. This suggests that the optimal number of clusters in this problem is four.

In Figs. 3(e,f) and 5 we display ε_t, and example PDF pairs (p_τ^k, p_τ^{Mk}) for $\epsilon \in \{0.1, 1\}$ and representative values of the forecast lead-time $t \in \{0, 0.02, 0.09\}$. Broadly speaking, ε_t has relatively small value at $t = 0$, but, because the dynamics of the reduced model differ systematically from those of the three-mode model, that value rapidly increases with t, until it reaches a maximum. At late times, ε_t decays to a K-independent equilibrium ε_{eq}. According to the equilibrium consistency condition in Eq. (24) is required to be small for skillful long-range forecasts.

As expected, ε_{eq} is an increasing function of ϵ. In the results of Fig. 3 we have $\varepsilon_{\text{eq}} = 0.008$ and 0.39, respectively for $\epsilon = 0.1$ and 1. That is, the $\epsilon = 0.1$ scalar reduced model is able to reproduce the equilibrium statistics of x accurately, but clearly the $\epsilon = 1$ reduced scalar model fails to be equilibrium consistent.

In all cases reported here, the maximum value of ε_t increases with the cluster number K. As illustrated in Fig. 5 the primary source of discrepancy is in the clusters containing large and positive values of x. The time-dependent PDFs conditioned on these clusters exhibit a significantly larger discrepancy as they relax to equilibrium compared to the clusters associated with small x, especially when ϵ is large.

4. Short- and medium-range forecasts in a nonstationary autoregressive model

We now relax the stationarity assumption of Sec. 3, and study predictability in stochastic dynamical systems with time-periodic equilibrium statistics. Such dynamical systems arise naturally in applications where seasonal effects are important; e.g., in AOS [23, 42, 72] and econometrics [34]. Here, a major challenge is to make high-fidelity forecasts given very short and noisy training time series [23]. A traditional, purely data-driven, approach to model-building in this context is to treat any time-dependent processes that are thought to be driving the observed time-periodic behavior as external factors, which are linearly coupled to a stationary autoregressive model of the dynamics. This leads to the so-called autoregressive factor models (ARX) [34], which are used widely in the aforementioned geophysical and financial applications.

Recently, Horenko [23] has developed an extension of the standard ARX methodology, in which the stationary ARX description is replaced by a convex combination of K locally-stationary ARX models. A key advantage of this approach is that it allows for distinct autoregressive dynamics to operate at a given time, depending on the affiliation of the system to one of K locally stationary models.

In this section, following a brief review of the nonstationary ARX formulation in Sec. 4.1, we apply the information-theoretic framework of Sec. 2 to assess the performance of nonstationary ARX models relative to the perfect model and globally-stationary ARX models. Throughout, we consider that the perfect model is a periodically-forced variant of the nonlinear scalar model in Eq. (28) with the parameter values listed in the $\epsilon = 0.1$ row of Table 1. Because of the presence of the quadratic and cubic nonlinearities and (more importantly) multiplicative noise, this is a particularly challenging application for both of the globally-stationary and nonstationary variants of ARX models. Thus, it should come as no surprise that in Sec. 4.3 we observe significant errors relative to the perfect model, especially when the effects of multiplicative noise are strong. Nevertheless, we find that the nonstationary ARX models can significantly outperform their globally-stationary counterparts, at least in the fidelity of time-dependent equilibrium statistics for these short training time series.

4.1. Constructing nonstationary autoregressive models via finite-element clustering

In the nonstationary ARX formalism [23], the true signal $x(t)$ in Eq. (1) (assumed here scalar for simplicity) is approximated by a system [23] of the form

$$x(t) = \sum_{k=1}^{K} \gamma_k(t) \left(\mu_k + \sum_{i=1}^{q} A_{ki} x(t - i\,\delta t) + B_k u(t) + C_k \epsilon(t) \right), \tag{32}$$

In the above, μ_k are model means; δt is a uniform sampling interval; $A_{k1}, \ldots, A_{kq}$ are autoregressive coefficients with memory depth q; B_k are couplings to the external factor $u(t)$; $\epsilon(t)$ is a Gaussian noise process with zero expectation and unit variance; and C_k are parameters coupling the noise to the observed time series. Moreover, $\gamma_k(t)$ are

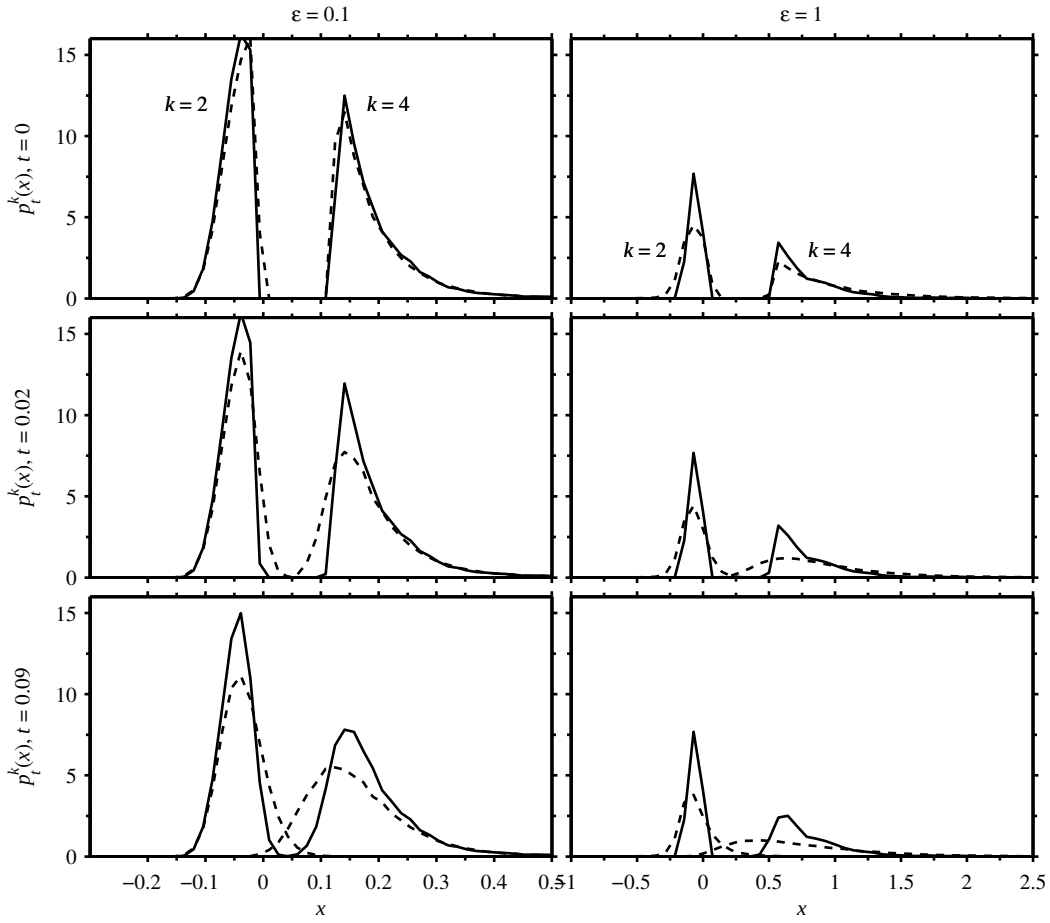

Figure 5: Time-dependent prediction probabilities for mode x in the perfect model [the three-mode model in Eq. (26)] and the imperfect model [the reduced scalar model in Eq. (28)] for $\epsilon = 0.1$ and $\epsilon = 1$. Plotted here in solid lines are the cluster-conditional PDFs $p_t^k(x)$ in the perfect model from Eq. (17) for clusters $k = 1$ and 4, ordered in order of increasing cluster coordinate θ_k in Eq. (15). The corresponding PDFs in the imperfect model, $p_t^{Mk}(x)$ from Eq. (22), are plotted in dashed lines. The forecast lead time t increases from top to bottom. As manifested by the discrepancy between p_t^k and p_t^{Mk}, the error in the imperfect model is significantly higher for $\epsilon = 1$ than 0.1. In both cases, a prominent source of error is that the scalar model relaxes to equilibrium at a faster rate than the true model, in the sense that the width of p_t^{Mk} increases more rapidly than the width of p_t^k (see also the correlation functions in Fig. 2). Moreover, the error in the imperfect models is more significant for large and positive values of x at the tails of the distributions in Fig. 1.

model weights satisfying the convexity conditions

$$\gamma_k(t) \geq 0 \quad \text{and} \quad \sum_{k=1}^{K} \gamma_k(t) = 1 \quad \text{for all } t. \tag{33}$$

In principle, given a training time series consisting of s samples of $x(t)$ in Eq. (14), the parameters $\theta_k = \{\mu_k, A_k, B_k, C_k\}$ for each model and the model weights in Eq. (33) are to be determined by minimizing the error functional

$$L(\Theta, \Gamma) = \sum_{k=1}^{K} \sum_{i=1}^{s} g(x(t - (i-1)\,\delta t), \theta_k), \tag{34}$$

with

$$g(x(t), \theta_k) = \left\| x(t) - \mu_k - \sum_{i=1}^{q} A_{ki} x(t - i\,\delta t) - B_k u(t) \right\|^2,$$
$$\Theta = \{\theta_1, \ldots, \theta_K\}, \quad \text{and} \quad \Gamma = \{\gamma_1(t), \ldots, \gamma_K(t)\}. \tag{35}$$

In practice, however, direct minimization of $L(\Theta, \Gamma)$ in Eq. (34) is generally an ill-posed problem [22–24], because of (i) non-uniqueness of $\{\Theta, \Gamma\}$ [due to the freedom in choosing $\gamma_k(t)$]; or (ii) lack of regularity of the model weights in Eq. (33) as a function of time, resulting in high-frequency, unphysical oscillations in $\gamma_k(t)$.

As demonstrated in Refs. [22–24], an effective strategy of dealing with the ill-posedness of the minimization of $L(\Theta, \Gamma)$ is to restrict the model weights $\gamma_k(t)$ to lie in a function space of sufficient regularity, such as the Sobolev space $W_{1,2}((0,T))$, or the space of functions of bounded variation $\mathrm{BV}((0,T))$ [73]. Here, we adopt the latter choice, since BV functions include functions with well-behaved jumps, and thus are suitable for describing sharp regime transitions.

As described in detail in Refs. [23, 28, 74], BV regularity may be enforced by augmenting the clustering minimization problem with a set of persistence constraints,

$$|\gamma_k|_{\mathrm{BV}} \leq C \quad \text{for all } k \in \{1, \ldots, K\}, \tag{36}$$

where

$$|\gamma_k|_{\mathrm{BV}} = \sum_{i=0}^{s-2} |\gamma_k(i\,\delta t) - \gamma_k((i-1)\,\delta t)|, \quad C \geq 0. \tag{37}$$

The above leads to a constrained linear optimization problem that can be solved by iteratively updating Θ and Γ. The special case with $K = 1$ reduces the problem to standard ARX models. In practical implementations of the scheme, the model affiliations $\gamma_k(t)$ are projected onto a suitable basis of finite element (FEM) basis functions [37], such as piecewise-constant functions. This reduces the number of degrees of freedom in the subspace of the optimization problem involving Γ, resulting in significant gains in computational efficiency.

In the applications below, we further require that the model affiliations are pure, i.e.,

$$\gamma_k(t) = \begin{cases} 1, & \text{if } k = S(t), \\ 0, & \text{otherwise,} \end{cases} \tag{38}$$

where

$$S(t) = \operatorname*{argmin}_j g(x(t), \theta_j). \tag{39}$$

This assumption is not necessary in general, but it facilitates the interpretation of results and time-integration of $x(t)$ in Eq. (32). Under the condition in Eq. (38), the BV seminorm in (36) measures the number of jumps in $\gamma_k(t)$. Thus, persistence in the BV sense here corresponds to placing an upper bound C on the number of jumps in the affiliation functions.

4.2. Making predictions in a time-periodic environment

In order to make predictions in the nonstationary ARX formalism, one must first advance the affiliation functions $\gamma_k(t)$ in Eq. (33) to times beyond the training time interval. One way of doing this is to construct a Markov model for the affiliation functions in Eq. (38) by fitting a K-state Markov generator matrix to the switching process Γ determined in the clustering optimization problem [23, 24], possibly incorporating time-dependent statistics associated with external factors [23]. However, this requires the availability of sufficiently-long training data to ensure convergence of the employed Markov generator algorithm [28, 75–77]. Because our objective here is to make predictions using very short training time series [23], we have opted to follow an alternative simple procedure, which directly exploits the time-periodicity in our applications of interest as follows.

Assume that the external factor $u(t)$ in Eq. (32) has period $\mathcal{T}$, and that the length $T = (s-1)\,\delta t$ of the training time series in Eq. (14) is at least $\mathcal{T}$. Then, for $t \geq T$, determine $\gamma_k(t)$ by periodic replication of $\gamma_k(t')$ with $t' \in [T - \mathcal{T}, T]$. This provides a mechanism for creating realizations of Eq. (32) given the value $x_0 = x(T)$ at the end of the training time series, leading in turn to a forecast probability distribution for x in the ARX model,

$$p_t^{Mx_0} = p(x(t) \mid x_0), \tag{40}$$

with $x(t)$ given by Eq. (32). The information theoretic error measures of Sec. 2 can then be computed by evaluating the entropy of the forecast distribution $p_t^{x_0}$ in the perfect model relative to the model distribution in Eq. (40). Note that, in accordance with Eq. (7) and Ref. [42], predictability in the perfect model is measured here relative to its time-dependent equilibrium measure and not relative to the (time-independent) distribution of period-averages of x.

4.3. Results and discussion

We consider that the true signal from nature (the perfect model) is given by the nonlinear scalar system in Eq. (28), forced with a periodic forcing of the form $F(t) = F_0 \cos(2\pi t/\mathcal{T} + \phi)$ of amplitude $F_0 = 0.5$, period $\mathcal{T} = 5$, and phase $\phi = 3\pi/4$ or $\pi/4$. As mentioned earlier, we adopt the parameter values in the row of Table 1 with $\epsilon = 0.1$. As illustrated in Figs. 6(a) and 7(a), with this choice of forcing and parameter values, the equilibrium PDF $p_t^{\rm eq}$ of the model is characterized by smooth transitions between low-variance, small-skewness phases when $F(t)$ is large and negative and high-variance positive-skewness phases when $F(t)$ is large and positive. The skewness of the distributions is a direct consequence of the multiplicative nature of the noise parameter β in Eq. (28), and poses a particularly high challenge for the ARX models in Eq. (32), where noise is additive and Gaussian.

We built stationary and nonstationary ARX models using as training data realizations of the perfect model of length $T = 2\mathcal{T}$, sampled uniformly every $\delta t = 0.01$ units (i.e., the total number of samples is $s = 1000$). To evaluate the nonstationary models we reduced the dimensionality of the $\gamma_k(t)$ affiliation functions by projecting them to an FEM basis consisting of $m = 200$ piecewise-constant functions of uniform width $\delta t_{\rm FEM} = T'/l = 5\,\delta t$. We solved the optimization problem in Eqs. (34)–(38) for $K \in \{2, 3\}$, systematically increasing the persistence parameter C from 1 to 40. In each case, we repeated the iterative optimization procedure 400 times, initializing (when possible) the first iteration with the solution determined at the previous value of C and the remaining 399 iterations with random initial data. The parameters m and C do not enter in the evaluation of the stationary models, since in that case the model parameters Θ can be determined analytically [23].

Following the method outlined in Sec. 4.2, we evaluated the ARX prediction probabilities $p_t^{Mx_0}$ in Eq. (40) up to lead time $t = T$ by replicating the model affiliation functions $\gamma_k(t)$ determined in the final portion of the training series with length T, and bin-counting realizations of $\tilde{x}(t)$ in Eq. (32) conditioned on the value at the end of the training time series. In the calculations reported here the initial conditions are $x_0 = 0.41$ and $x_0 = -0.098$, respectively for $\phi = \pi/4$ and $3\pi/4$. To estimate $p_t^{Mx_0}$, we nominally used $r = 1.2 \times 10^7$ realizations of $x(t)$ in the scalar and ARX models, which we binned over $b = 50$ uniform bins in the interval $[-0.5, 0.6]$. The same procedure was used to estimate the finite-time and equilibrium prediction probabilities in the perfect model, $p_t^{x_0}$ and $p_t^{\rm eq}$, respectively. All relative-entropy calculations required to evaluate the skill and error metrics of Sec. 2 ($\mathcal{D}_t^{x_0}$ and $\mathcal{E}_t^{x_0}$) were then carried out using the standard trapezoidal rule with the histograms for $p_t^{x_0}$, $p_t^{Mx_0}$, and $p_t^{\rm eq}$. We checked for robustness of our entropy calculations by halving r and/or b. Neither of these imparted significant changes on our results.

In separate calculations, we have studied nonstationary ARX models where, instead of a periodic continuation of the model affiliation sequence fitted in the training data, a nonstationary K-state Markov process was employed to evolve the integer-valued affiliation function $S(t)$ dynamically. Here, to incorporate the effects of the external forcing in the switching process, the Markov process was constructed by fitting a transition matrix of the form $P(t) = P_0 + P_1 F(t)$ in the $S(t)$ sequence obtained in the training stage [28]. However, the small number of jumps in the training data precluded a reliable estimation of P_0 and P_1, resulting in no improvement of skill compared to models based on periodic continuation of $S(t)$.

Hereafter, we restrict attention to nonstationary ARX models with $K = 3$ and $C = 8$, and their stationary ($K = 1$) counterparts. These models, displayed in Table 3 and Figs. 6–8, exhibit the representative types of behavior that are of interest to us here, and are also robust with respect to changes in C and/or the number of FEMs.

To begin, note an important qualitative difference between the systems with forcing phase $\phi = 3\pi/4$ and $\pi/4$, which can be seen in Figs. 6,7(a): The variance of the $\phi = \pi/4$ system at the beginning of the prediction period is significantly higher than the corresponding variance observed for $\phi = 3\pi/4$. As a result, the perfect-model predictability, as measured by the δ_t skill score from Eq. (9), drops more rapidly in the former model. In both cases, however, predictability beyond the time-periodic equilibrium becomes negligible beyond $t \simeq 1.5$ time units, or $0.3\mathcal{T}$, as manifested by the small value of the δ_t skill score in Figs. 6,7(e). Thus, even though predictions in the model with $\phi = \pi/4$ are inherently less skillful at early times than in the $\phi = 3\pi/4$ model, the best that one can expect in either model of forecasts with lead times beyond about $t = 1.5$ is to reproduce the equilibrium statistics with high fidelity. Given the short length of the training series this is a challenging problem for any predictive model, including the stationary and nonstationary ARX models employed here.

A second key point is that all models in Table 3 have the property

$$|A_k| < 1 \quad \text{for all } k \in [1, \ldots, K], \tag{41}$$

which here is sufficient to guarantee the existence of a time-periodic statistical equilibrium state. The existence of a statistical equilibrium state is a property of many complex dynamical systems arising in applications. Therefore, if one is interested in making predictions over lead times approaching or exceeding the equilibration time of the perfect model, it is natural to require at a minimum that the ARX models have a well-behaved equilibrium distribution $p_t^{M,\rm eq}$ [the imperfect-model analog of Eq. (3)]. In the globally-stationary ARX models studied here, Eq. (41) is also a necessary condition for the existence of $p_t^{M,\rm eq}$. On the other hand, nonstationary ARX models can contain locally-stationary unstable components (i.e., some autoregressive couplings with $|A_k| > 1$), and remain bounded in equilibrium. As has been noted elsewhere [78], high fi-

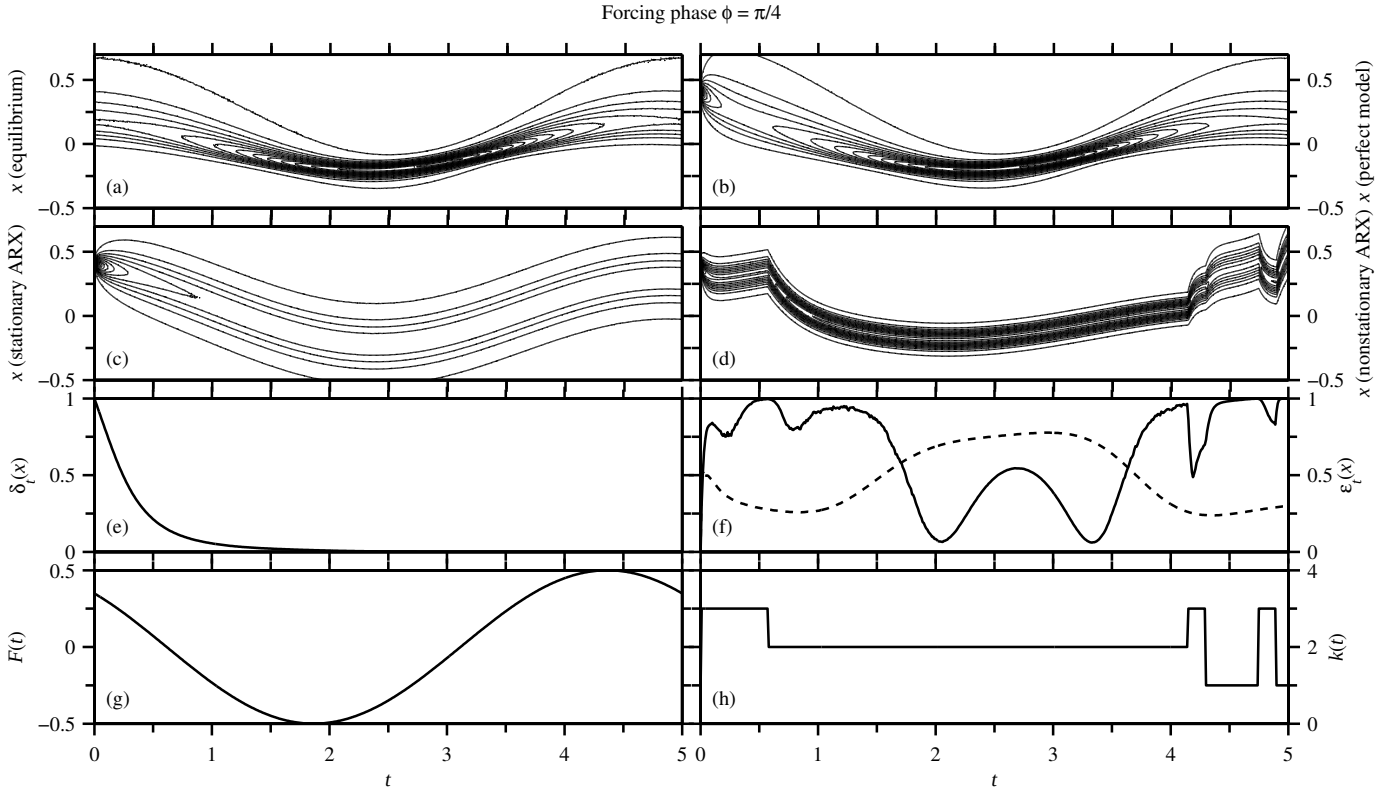

Figure 6: Time-dependent PDFs, predictive skill in the perfect model, and ARX model error for the system in Table 3 with forcing phase $\phi = \pi/4$. Shown here are (a) contours of the equilibrium distribution p_t^{eq} of mode x in the true model as a function of x and time; (b) contours of the time-dependent PDF $p_t^{x_0}(x)$ in the perfect model, conditioned on initial data $x_0 = 0.41$; (c,d) contours of the time-dependent PDF $p_t^{Mx_0}$ in the globally-stationary and nonstationary ARX models ($K = 3$); (e) the predictive skill score δ_t in the perfect model; (f) the normalized error ε_t in the ARX models; (f) the time-periodic forcing $F(t)$; (g) the cluster-affiliation sequence $k(t)$ in the nonstationary ARX model, determined by replicating the portion of the affiliation sequence in the training time series with $t \in [\mathcal{T}, 2\mathcal{T}]$ (see Fig. 8). The contour levels in Panels (a)–(d) span the interval $[0.1, 15]$, and are spaced by 0.92.

Table 3: Properties of non-stationary ($K = 3$) and stationary ARX models of the nonlinear scalar stochastic with time-periodic forcing.

	$\phi = \pi/4$				$\phi = 3\pi/4$			
State	μ_k	A_k	σ_k	B_k	μ_k	A_k	σ_k	B_k
1	0.1568	0.8721	0.0370	-0.2204	0.0583	0.7710	0.0230	0.0269
2	-0.0022	0.9581	0.0122	0.0115	-0.0021	0.9672	0.0120	0.0117
3	0.0607	0.8327	0.0326	-0.0444	0.0527	0.7165	0.0205	-0.0198
Stationary	6×10^-4	0.9836	0.0217	0.0107	-5×10^{-4}	0.9785	0.0150	0.0106

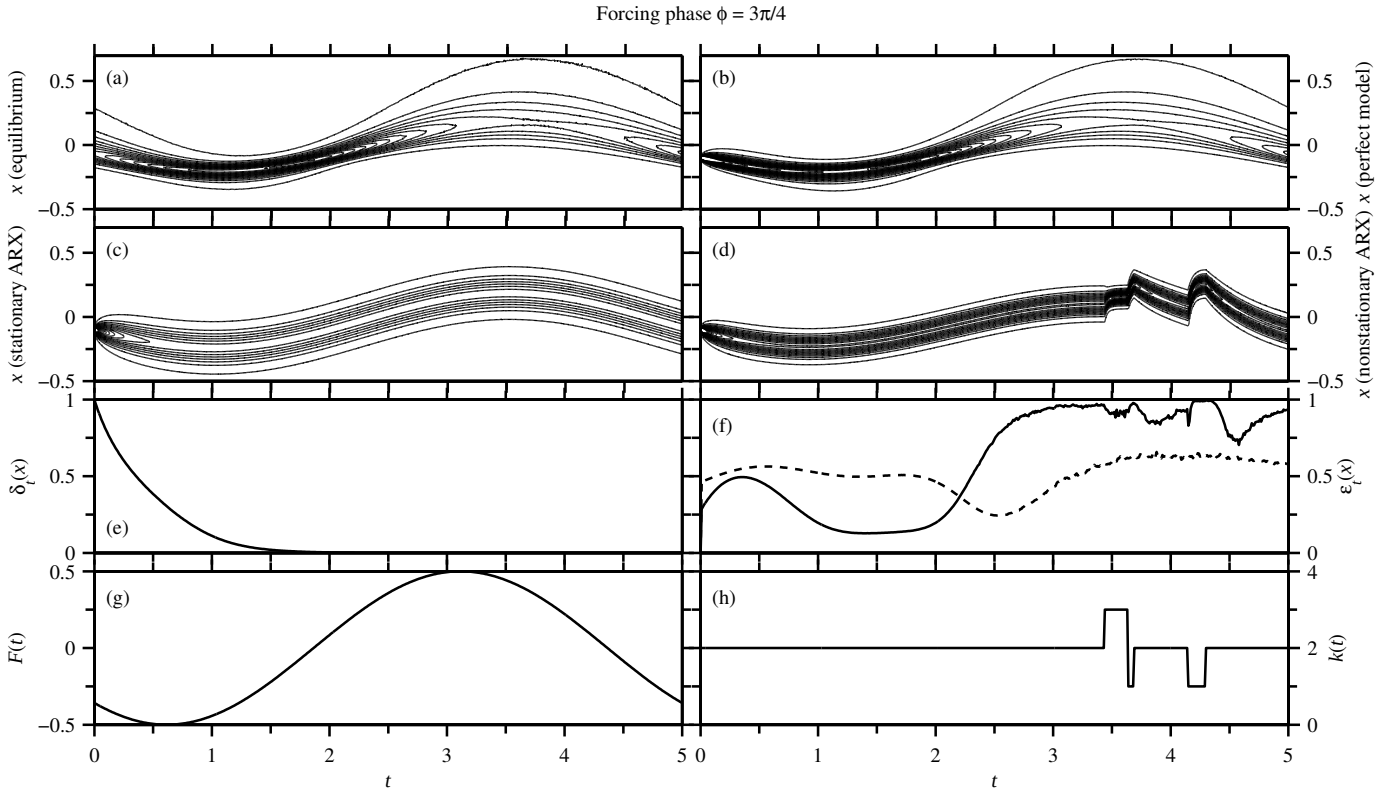

Figure 7: Time-dependent PDFs, predictive skill in the perfect model, and ARX model error for the system in Table 3 with forcing phase $\phi = 3\pi/4$. Shown here are (a) contours of the equilibrium distribution p_t^{eq} of mode x in the true model as a function of x and time; (b) contours of the time-dependent PDF $p_t^{x_0}(x)$ in the perfect model, conditioned on initial data $x_0 = -0.098$; (c,d) contours of the time-dependent PDF $p_t^{Mx_0}$ in the globally-stationary and nonstationary ARX models ($K = 3$); (e) the predictive skill score δ_t in the perfect model; (f) the normalized error ε_t in the ARX models; (f) the time-periodic forcing $F(t)$; (g) the cluster-affiliation sequence $k(t)$ in the nonstationary ARX model, determined by replicating the portion of the affiliation sequence in the training time series with $t \in [\mathcal{T}, 2\mathcal{T}]$ (see Fig. 8). The contour levels in Panels (a)–(d) span the interval $[0.1, 15]$ and are spaced by 0.92.

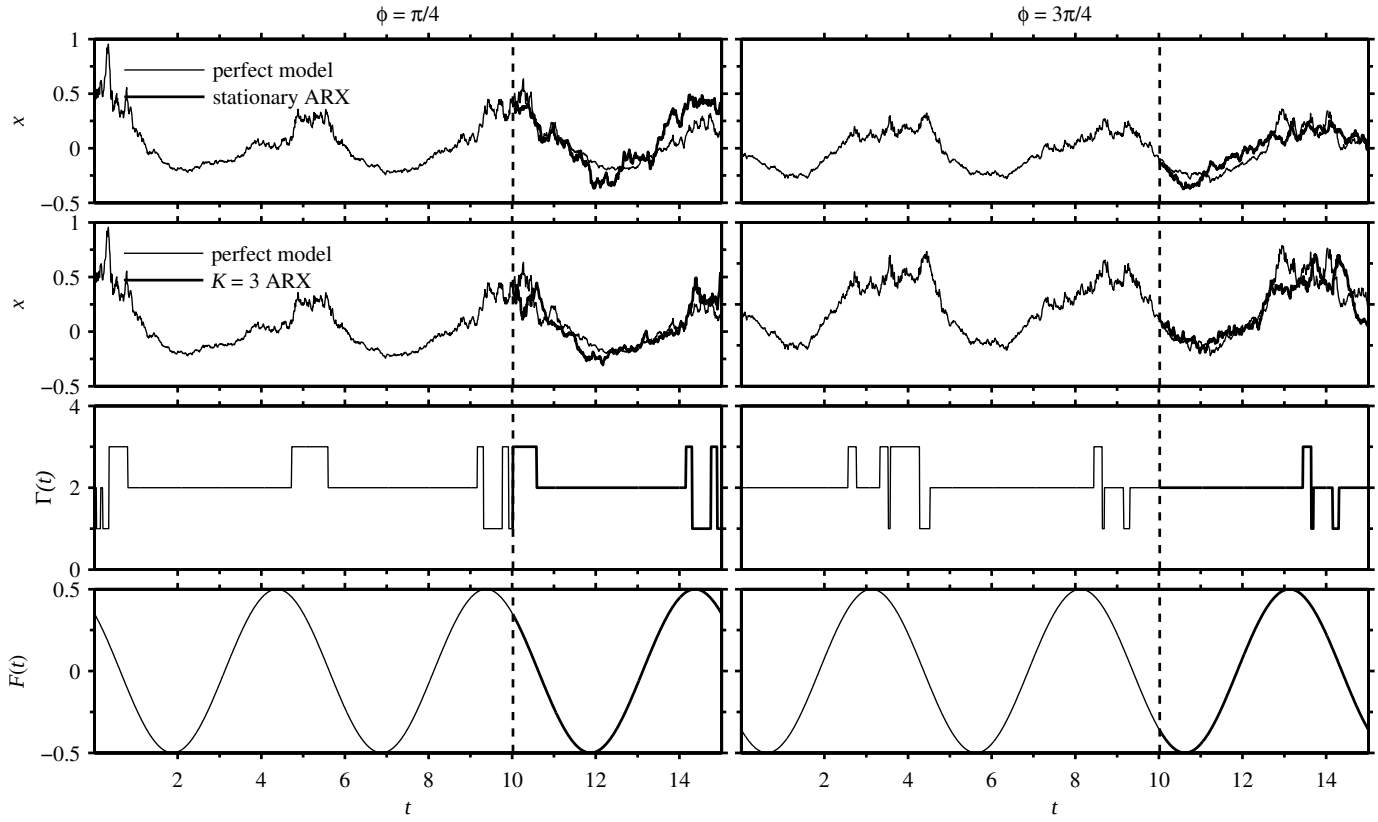

Figure 8: Training ($t \in [0, 10]$ and prediction stages ($t \in [10, 15]$) of globally stationary and nonstationary ARX models. The panels in the first two rows display in thick solid lines realizations of globally stationary and nonstationary ($K = 3$) ARX models, together with a sample trajectory of the true model (thin solid lines). Also shown are the cluster-affiliation sequence $k(t) = \text{argmax}_j \, \gamma_j(t)$ of the stationary ARX models and the external periodic forcing $F(t)$ (here, the forcing period is $\mathcal{T} = 5$. The parameters of these models are listed in Table 3.

delity and/or skill can exist despite structural instability of this type.

We now discuss model fidelity, first in the context of globally stationary ARX models. As shown in Figs. 6,7(c), these models tend to overestimate the variance of the true model during periods of negative forcing. Evidently, these "$K = 1$" models do not have sufficient flexibility to accommodate the changes in variance due to multiplicative noise in the true model. These ARX models also fail to reproduce the skewness towards large and positive x values in the true model, but this deficiency is shared in common with the locally-stationary models, due to the structure of the noise term in Eq. (32).

Consider now the nonstationary ARX models with $K = 3$. As expected intuitively, in both of the $\phi = 3\pi/4$ and $\pi/4$ cases the model-affiliation function is constant in low-variance periods, and switches more frequently in high-variance periods [see Figs. 6,7(h)]. Here, a prominent aspect of behavior is that periods of high variance in the true model are replaced by rapid transitions between locally-stationary models, which generally underestimate the variance in the true model. For this reason, the model error ε_t in the non-stationary ARX models generally exceeds the error in the stationary models in these regimes. In the system with $\phi = 3\pi/4$ this occurs at late times [$t \gtrsim 2.2$ in Fig. 7(f)], but the error is large for both early and late times, $t \in [0, 1.5] \cup [3.5, 5]$, in the more challenging case with $\phi = \pi/4$ in Fig. 6(f).

The main strength, however, of the nonstationary ARX models is that they are able to predict with significantly higher fidelity during low-variance periods in the true model. The improvement in performance is especially noticeable in the system with $\phi = 3\pi/4$, where the $K = 3$ model outperforms the globally stationary ARX model at early times ($t \lesssim 1.5$), as well as in the the regime with $t \in [1.5, 2.5]$, where no significant predictability exists beyond the time-periodic equilibrium measure. The fidelity of the non-stationary ARX model in reproducing the equilibrium statistics in this case is remarkable given that only two periods of the forcing were used as training data. In the example with $\phi = \pi/4$ the gain in fidelity is less impressive. Nevertheless, the $K = 3$ model significantly outperforms the globally-stationary model. In both $\phi = \pi/4$ and $3\pi/4$ cases the coupling B_k to the external factor is positive in the low-variance phase with $k = 2$ (see Table 3).

It therefore follows from this analysis that nonstationary models exploit the additional flexibility beyond the globally stationary models to preferentially bring down the value of the integrand in the clustering functional in Eq. (34) (i.e., the "error density") over certain sub-intervals of training series. This entails significant improvements to predictive fidelity over those subintervals. Intriguingly, the reduction of model error arises out of global optimization over the training time interval, i.e., through a non-causal process.

It is also interesting to note that the $K = 3$ models with high predictive fidelity would actually be ruled out if assessed by means of model discrimination analysis based on the AIC [53]. According to this criterion, the optimal model in a class of competing models is the one with the smallest value of

Table 4: The Akaike information criterion AIC from Eq. (42) for the models in Table 3

	AIC ($\phi = \pi/4$)	AIC ($\phi = 3\pi/4$)
$K = 3$	-1.204×10^4	-1.24×10^4
Stationary	-1.33×10^4	-1.48×10^4

$$\mathrm{AIC} = -2\mathcal{L} + 2\mathcal{N}, \tag{42}$$

where $\mathcal{L}$ is a log-likelihood function measuring the closeness of fit of the training data by the model, and $\mathcal{N}$ the number of free parameters in the model. Thus, AIC penalizes models that tend to overfit the data by employing unduly large numbers of parameters. Given parametric distributions ψ_k describing the residuals $r_k(t) = g(x(t), \theta_k)$ from Eq. (35) [the $r_k(t)$ are assumed to be statistically independent], the likelihood and penalty components of the AIC functional for the nonstationary ARX models in Eq. (32) are [28]

$$\begin{aligned} \mathcal{L} &= \sum_{i=1}^{s} \log\left(\sum_{k=1}^{K} \gamma_k((i-1)\,\delta t)\psi_k(r_k((i-1)\,\delta t)) \right), \\ \mathcal{N} &= KN_{\mathrm{ARX}} + \sum_{k=1}^{K-1} |\gamma_k|_{\mathrm{BV}}, \end{aligned} \tag{43}$$

with N_{ARX} the number of parameters in each locally stationary ARX model ($N_{\mathrm{ARX}} = 3$ for μ_k, A_k, B_k; the σ_k noise intensity is determined using the latter three parameters [23]), and $|\gamma_k|_{\mathrm{BV}}$ the number of jumps in $\gamma_k(t)$ [see Eq. (37)].

Here, we set ψ_k to the exponential distribution, $\psi_k(r) = \lambda_k e^{-\lambda_k r}$, with λ_k determined empirically from the mean of $r_k(t)$. The exponential distribution yielded higher values of log-likelihood than the χ^2 distribution for our datasets, and also has an intuitive interpretation as the least biased (maximum entropy) distribution given the observed λ_k. According to the AIC functional in Eq. (42), whose values are listed in Table 4, the globally stationary models are favored over their nonstationary counterparts for both values of the external-forcing phase ϕ considered here. Thus, the optimal models in the sense of AIC are not necessarily the highest-performing models in the sense of predictive fidelity.

5. Conclusions

In this paper we have developed information-theoretic strategies to quantify predictive skill and assess the fidelity of predictions with imperfect models in (i) long-range, coarse grained forecasts in complex nonlinear systems; (ii) short- and medium-range forecasts in systems with time-periodic external forcing. We have demonstrated these

strategies using instructive prototype models, which are of widespread applicability in applied mathematics, physical sciences, engineering, and social sciences.

Using as an example a three-mode stochastic model with dyad interactions, observed through a scalar slow mode carrying about 0.5% of the total variance, we demonstrated that suitable coarse-grained partitions of the set of initial data reveal long-range predictability, and provided a clustering-algorithm to evaluate these partitions from ergodic trajectories in equilibrium. This algorithm requires no detailed treatment of initial data and does not impose parametric forms on the probability distributions for ensemble forecasts. As a result, objective measures of predictability based on relative entropy can be evaluated practically in this framework.

The same information-theoretic framework can be used to quantify objectively the error in imperfect models; an issue of strong contemporary interest in science and engineering. Here, we have put forward a scheme which assesses the skill of imperfect models based on three relative-entropy metrics: (i) the lack of information (or ignorance) $\mathcal{E}_{\rm eq}$ of the imperfect model in equilibrium; (ii) the lack of information $\mathcal{E}_t$ during model relaxation from equilibrium; (iii) the discrepancy of prediction distributions $\mathcal{D}_t^M$ in the imperfect model relative to its equilibrium. In this scheme $\mathcal{E}_{\rm eq} \ll 1$, is a necessary, but not sufficient, condition for long-range forecasting skill. If a model meets that condition (called here equilibrium consistency) and the analogous condition at finite lead times, $\mathcal{E}_t \ll 1$ then $\mathcal{D}_t^M$ is a meaningful measure of predictive skill. Otherwise, $\mathcal{D}_t^M$ conveys false skill. We have illustrated how this scheme works in an application where the three-mode dyad model is treated as the true model, and the role of imperfect model is played by a cubic scalar stochastic model with multiplicative noise (which is formally accurate in the limit of infinite timescale separation between the slow and fast modes).

In the context of time-periodic models our analysis has revealed that recently proposed nonstationary autoregressive models [23], based on bounded-variation finite-element clustering can significantly outperform their stationary counterparts in the fidelity of short- and medium-range predictions in challenging nonlinear systems with multiplicative noise. In particular, we found high fidelity in a three-state autoregressive model at short times and in reproducing the equilibrium statistics at later lead times, despite the fact that only two periods of the forcing were used as training data.

In future work we plan to extend the nonstationary ARX formalism to explicitly incorporate physically-motivated nonlinearities in the autoregressive model.

Acknowledgments

This research of Andrew Majda is partially supported by NSF grant DMS-0456713, from ONR DRI grants N25-74200-F6607 and N00014-10-1-0554, and from DARPA grants N00014-07-10750 and N00014-08-1-1080. Dimitrios Giannakis is supported as a postdoctoral fellow through the last three agencies. The authors wish to thank Paul Fischer for providing computational resources at Argonne National Laboratory. Much of this research was developed while the authors were participants in the long program at the Institute for Pure and Applied Mathematics (IPAM) on Hierarchies for Climate Science, which is supported by NSF, and in a recent month-long visit of DG and AJM to the University of Lugano.

References

[1] E. N. Lorenz, Deterministic nonperiodic flow, J. Atmos. Sci. 20 (1963) 130–141.
[2] E. N. Lorenz, A study of predictability of a 28-variable atmospheric model, Tellus 17 (1968) 321–333.
[3] E. N. Lorenz, The predictability of a flow which possesses many scales of moition, Tellus 21 (1969) 289–307.
[4] E. S. Epstein, Stochastic dynamic predictions, Tellus 21 (1969) 739–759.
[5] D. Ruelle, F. Takens, On the nature of turbulence, Commun. Math. Phys. 20 (1971) 167–192.
[6] J. A. Vastano, H. L. Swinney, Information transport in spatiotemporal systems, Phys. Rev. Lett. 60 (1988) 1773.
[7] K. Sobczyk, Information dynamics: Premises, challenges and results, Mech. Syst. Signal Pr. 15 (2001) 475–498.
[8] A. J. Majda, J. Harlim, Information flow between subspaces of complex dynamical systems, Proc. Natl. Acad. Sci. 104 (2007) 9558–9563.
[9] M. A. Katsoulakis, A. J. Majda, D. G. Vlachos, Coarse-grained stochastic processes and monte carlo simulations in lattice systems, J. Comput. Phys. 186 (2003) 250–278.
[10] M. A. Katsoulakis, D. G. Vlachos, Coarse-grained stochastic processes and kinetic Monte Carlo simulators for the diffusion of interacting particles, J. Chem. Phys. 119 (2003) 9412–9427.
[11] M. A. Katsoulakis, A. J. Majda, A. Sopasakis, Intermittency, metastability and coarse-graining for coupled deterministic–stochastic lattice systems, Nonlinearity 19 (2006) 1021–1047.
[12] M. A. Katsoulakis, P. Plecháč, A. Sopasakis, Error analysis of coarse-graining for stochastic lattice dynamics, J. Numer. Anal. 44 (2006) 2270–2296.
[13] A. Chatterjee, D. G. Vlachos, An overview of spatial microscopic and accelerated kinetic Monte Carlo methods, J. Computer-Aided Mater. Des. 14 (2007) 253–308.
[14] L.-Y. Leung, G. R. North, Information theory and climate prediction, J. Climate 3 (1990) 5–14.
[15] T. Schneider, S. M. Griffies, A conceptual framework for predictability studies, J. Climate 12 (1999) 3133–3155.
[16] R. Kleeman, Measuring dynamical prediction utility using relative entropy, J. Atmos. Sci. 59 (2002) 2057–2072.
[17] M. Roulston, L. Smith, Evaluating probabilistic forecasts using information theory, Mon. Weather Rev. 130 (2002) 1653–1660.
[18] T. DelSole, Predictability and information theory. Part I: Measures of predictability, J. Atmos. Sci. 61 (2004) 2425–2440.
[19] T. DelSole, Predictability and information theory. Part II: Imperfect models, J. Atmos. Sci. 62 (2005) 3368–3381.
[20] C. Franzke, D. Crommelin, A. Fischer, A. J. Majda, A hidden Markov model perspective on regimes and metastability in atmospheric flows, J. Climate 21 (2008) 1740–1757.
[21] C. Franzke, I. Horenko, A. J. Majda, R. Klein, Systematic metastable regime identification in an AGCM, J. Atmos. Sci. 66 (2009) 1997–2012.
[22] I. Horenko, On robust estimation of low-frequency variability trends in discrete Markovian sequences of atmospheric circulation patterns, J. Atmos. Sci. 66 (2009) 2059–2072.
[23] I. Horenko, On the identification of nonstationary factor models

and their application to atmospheric data analysis, J. Atmos. Sci. 67 (2010) 1559–1574.
[24] I. Horenko, On clustering of non-stationary meteorological time series, Dyn. Atmos. Oceans 49 (2010) 164–187.
[25] H. Teng, G. Branstator, Initial-value predictability of prominent modes of North Pacific subsurface temperature in a CGCM, Climate Dyn. (2010). In press.
[26] D. Giannakis, A. J. Majda, Quantifying the predictive skill in long-range forecasting. Part I: Coarse-grained predictions in a simple ocean model, J. Climate (2011). Submitted.
[27] D. Giannakis, A. J. Majda, Quantifying the predictive skill in long-range forecasting. Part II: Model error in coarse-grained Markov models with application to ocean-circulation regimes, J. Climate (2011). Submitted.
[28] I. Horenko, Nonstationarity in multifactor models of discrete jump processes, memory and application to cloud modeling, J. Atmos. Sci. (2011). Early online release.
[29] P. Deuflhard, M. Dellnitz, O. Junge, C. Schütte, Computation of essential molecular dynamics by subdivision techniques I: basic concept, Lect. Notes Comp. Sci. Eng. 4 (1999) 98.
[30] P. Deuflhard, W. Huisinga, A. Fischer, C. Schütte, Identification of almost invariant aggregates in reversible nearly uncoupled markov chains, Linear Alg. Appl. 315 (2000) 39.
[31] F. Noé, J. M. Smith, Transition networks: A unifying theme for molecular simulation and computer science, in: Mathematical Modeling of Biological Systems, Volume I, Modeling and Simulation in Science, Engineering and Technology, Birkhäuser, Boston, 2007, pp. 121–137.
[32] S. Haider, G. N. Parkinson, S. Neidle, Molecular dynamics and principal components analysis of human telomeric quadruplex multimers, Biophys. J. 95 (2008) 296–311.
[33] I. Horenko, C. Schütte, On metastable conformation analysis of nonequilibrium biomolecular time series, Multiscale Model. Simul. 8 (2010) 701–716.
[34] R. S. Tsay, Analysis of Financial Time Series, Wiley, Hoboken, 2010.
[35] L. Putzig, D. Becherer, I. Horenko, Optimization of a futures portfolio utilizing numerical market phase-detection, SIAM J. Finan. Math. 1 (2010) 752–779.
[36] I. Horenko, On simultaneous data-based dimension reduction and hidden phase identification, J. Atmos. Sci. 65 (2008) 1941–1954.
[37] I. Horenko, Finite element approach to clustering of multidimensional time series, SIAM J. Sci. Comput. 32 (2010) 62–83.
[38] J. de Wiljes, I. Majda, A. J. ahd Horenko, An adaptive Markov chain Monte Carlo approach to time series clustering with regime transition behavior, J. Comput. Phys. (2010). Submitted.
[39] J. Berner, G. Branstator, Linear and nonlinear signatures in planetary wave dynamics of an AGCM: Probability density functions, J. Atmos. Sci. 64 (2007) 117–136.
[40] A. J. Majda, C. Franzke, A. Fischer, D. T. Crommelin, Distinct metastable atmospheric regimes despite nearly Gaussian statistics: A paradigm model, Proc. Natl. Acad. Sci. 103 (2006) 8309–8314.
[41] C. Franzke, A. J. Majda, G. Branstator, The origin of nonlinear signatures of planetary wave dynamics: Mean phase space tendencies and contributions from non-Gaussianity, J. Atmos. Sci. 64 (2007) 3988.
[42] A. J. Majda, X. Wang, Linear response theory for statistical ensembles in complex systems with time-periodic forcing, Comm. Math. Sci. 8 (2010) 145–172.
[43] A. J. Majda, B. Gershgorin, Quantifying uncertainty in climage change science through empirical information theory, Proc. Natl. Acad. Sci. 107 (2010) 14958–14963.
[44] C. E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423,623–656.
[45] A. J. Majda, R. Kleeman, D. Cai, A mathematical framework for predictability through relative entropy, Methods Appl. Anal. 9 (2002) 425–444.
[46] R. V. Abramov, A. J. Majda, R. Kleeman, Information theory and predictability for low-frequency variability, J. Atmos. Sci. 62 (2005) 65–87.
[47] A. J. Majda, R. V. Abramov, M. J. Grote, Information Theory and Stochastics for Multiscale Nonlinear Systems, volume 25 of *CRM Monograph Series*, Americal Mathematical Society, Providence, 2005.
[48] T. A. Cover, J. A. Thomas, Elements of Information Theory, Wiley-Interscience, Hoboken, 2 edition, 2006.
[49] T. DelSole, M. K. Tippett, Predictability: Recent insights from information theory, Rev. Geophys. 45 (2007) RG4002.
[50] R. Kleeman, Information theory and dynamical system predictability, in: Isaac Newton Institute Preprint Series, NI10063, 2010, pp. 1–33.
[51] A. J. Majda, B. Gershgorin, Y. Yuan, Low-frequency climate response and fluctuation–dissipation theorems: Theory and practice, J. Atmos. Sci. 67 (2010) 1186.
[52] A. J. Majda, I. I. Timofeyev, E. Vanden Eijnden, Systematic strategies for stochastic mode reduction in climate, J. Atmos. Sci. 60 (2003) 1705.
[53] H. Akaike, Information theory and an extension of the maximum likelihood principle, in: B. N. Petrov, F. Caski (Eds.), Proceedings of the Second International Symposium on Information Theory, Akademiai Kiado, Budapest, 1973, p. 610.
[54] A. D. R. McQuarrie, C.-L. Tsai, Regression and Time Series Model Selection, World Scientific, Singapore, 1998.
[55] H. Bozdogan, Akaike's information criterion and recent developments in information complexity, J. Math. Psychol. 44 (2000) 62.
[56] H. Joe, Relative entropy measures of multivariate dependence, J. Amer. Stat. Assoc. 84 (1989) 157–164.
[57] B. Khouider, A. St-Cyr, A. J. Majda, J. Tribbia, MJO and convectively coupled waves in a coarse resolution GCM with a simple multicloud parametrization, J. Atmos. Sci. (2011). Early online release.
[58] A. J. Majda, B. Gershgorin, Improving model fidelity and sensitivity for complex systems through empirical information theory, Proc. Natl. Acad. Sci. (2011). Submitted.
[59] D. Madigan, A. Raftery, C. Volinsky, J. Hoeting, Bayesian model averaging, in: Proceedings of the AAAI Workshop on Integrating Multiple Learned Models, Portland, OR, pp. 77–83.
[60] J. A. Hoeting, D. Madigan, A. E. Raftery, C. T. Volinsky, Bayesian model averaging: A tutorial, Stat. Sci. 14 (1999) 382–401.
[61] A. E. Raftery, T. Gneiting, F. Balabdaoui, M. Polakowski, Using bayesian model averaging to calibrate forecast ensembles, Mon. Weather Rev 133 (2005) 1155–1173.
[62] G. Branstator, J. Berner, Linear and nonlinear signatures in the planetary wave dynamics of an AGCM: Phase space tendencies, J. Atmos. Sci. 62 (2005) 1792–1811.
[63] G. A. Meehl, et al., Decadal prediction. can it be skillful?, Bull. Amer. Meteor. Soc. 90 (2009) 1467–1485.
[64] B. W. Silverman, Density Estimation for Statistics and Data Analysis, volume 26 of *Monographs on Statistics and Applied Probability*, Chapman & Hall/CRC, Boca Raton, 1986.
[65] J. MacQueen, Some methods for classification and analysis of multivariate observations, in: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, volume 1, University of California Press, Berkeley, 1967, pp. 281–287.
[66] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification, Wiley-Interscience, New York, 2 edition, 2000.
[67] T. DelSole, J. Shukla, Model fidelity versus skill in seasonal forecasting, J. Climate 23 (2010) 4794–4806.
[68] A. J. Majda, C. Franzke, D. Crommelin, Normal forms for reduced stochastic climate models, Proc. Natl. Acad. Sci. 106 (2009) 3649.
[69] G. F. Lawler, Introduction to Stochastic Processes, Chapman & Hall/CRC, Boca Raton, 2006.
[70] C. Gardiner, Stochastic Methods: A Handbook for the Natural and Social Sciences, Springer Series in Synergetics, Springer, Berlin, 4 edition, 2010.

[71] A. J. Majda, I. I. Timofeyev, E. Vanden Eijnden, Models for stochastic climate predicition, Proc. Natl. Acad. Sci. 96 (1999) 14687.
[72] B. Gershgorin, A. J. Majda, A test model for fluctuation–dissipation theorems with time-periodic statistics, Physica D 239 (2010) 1741–1757.
[73] L. Ambrosio, N. Fusco, D. Pallara, Functions of Bounded Variation and Free-Discontinuity Problems, Oxford University Press, 2000.
[74] I. Horenko, Parameter identification in nonstationary Markov chains with external impact and its application to computational sociology, SIAM J. Multiscale Model. Sim. (2011).
[75] D. T. Crommelin, E. Vanden-Eijnden, Fitting timeseries by continuous-time Markov chains: A quadratic programming approach, J. Comput. Phys. 217 (2006) 782–805.
[76] P. Metzner, E. Dittmer, T. Jahnke, C. Schütte, Generator estimation of Markov jump processes based on incomplete observations equidistant in time, J. Comput. Phys. 227 (2007) 353–375.
[77] P. Metzner, I. Horeko, C. Schütte, Generator estimation of Markov jump processes based on incomplete observations nonequidistant in time, Phys. Rev. E 76 (2007) 066702.
[78] A. J. Majda, R. Abramov, B. Gershgorin, High skill in low-frequency climate response through fluctuation dissipation theorems despite structural instability, Proc. Natl. Acad. Sci 107 (2010) 581–586.