SCALABLE BAYESIAN REDUCED-ORDER MODELS FOR SIMULATING HIGH-DIMENSIONAL MULTISCALE DYNAMICAL SYSTEMS *

PHAEDON-STELIOS KOUTSOURELAKIS † AND ELIAS BILIONIS ‡

Abstract. While existing mathematical descriptions can accurately account for phenomena at microscopic scales (e.g. molecular dynamics), these are often high-dimensional, stochastic and their applicability over macroscopic time scales of physical interest is computationally infeasible or impractical. In complex systems, with limited physical insight on the coherent behavior of their constituents, the only available information is data obtained from simulations of the trajectories of huge numbers of degrees of freedom over microscopic time scales. This paper discusses a Bayesian approach to deriving probabilistic coarse-grained models that simultaneously address the problems of identifying appropriate reduced coordinates and the effective dynamics in this lower-dimensional representation. At the core of the models proposed lie simple, low-dimensional dynamical systems which serve as the building blocks of the global model. These approximate the latent, generating sources and parameterize the reduced-order dynamics. We discuss parallelizable, online inference and learning algorithms that employ Sequential Monte Carlo samplers and scale linearly with the dimensionality of the observed dynamics. We propose a Bayesian adaptive time-integration scheme that utilizes probabilistic predictive estimates and enables rigorous concurrent s imulation over macroscopic time scales. The data-driven perspective advocated assimilates computational and experimental data and thus can materialize data-model fusion. It can deal with applications that lack a mathematical description and where only observational data is available. Furthermore, it makes non-intrusive use of existing computational models.

Key words.

AMS subject classifications.

1. Introduction.

The present paper is concerned with the development of probabilistic coarse-grained models for high-dimensional dynamical systems with a view of enabling multiscale simulation. We describe a unified treatment of complex problems described by large systems of deterministic or stochastic ODEs and/or large number of data streams. Such systems arise frequently in modern multi-physics applications either due to the discrete nature of the system (e.g. molecular dynamics) or due to discretization of spatiotemporal models (e.g. PDEs):

$$\frac{d\boldsymbol{y}_t}{dt} = \boldsymbol{f}(\boldsymbol{y}_t), \quad \boldsymbol{y} \in \mathcal{Y} \tag{1.1}$$

where $dim(\mathcal{Y}) >> 1$ (e.g. $\mathbb{R}^d$, $d >> 1$). Stochastic versions are also frequently encountered:

$$\frac{d\boldsymbol{y}_t}{dt} = \boldsymbol{f}(\boldsymbol{y}_t; \boldsymbol{u}_t) \tag{1.2}$$

where $\boldsymbol{u}_t$ is a driving stochastic process (i.e. Wiener process). Uncertainties could also appear in the initial conditions that accompany the aforementioned systems of equations.

*This work was supported by the OSD/AFOSR MURI'09 award to Cornell University on uncertainty quantification

†School of Civil and Environmental Enginerring & Center for Applied Mathematics, 369 Hollister Hall, Cornell University, Ithaca, NY 14853, (pk285@cornell.edu)

‡Center for Applied Mathematics, 464 Hollister Hall, Cornell University, Ithaca, NY 14853, (ib227@cornell.edu).

Even though the numerical solution of (stochastic) ODEs is a well-studied subject and pertinent computational libraries are quite mature, traditional schemes are impractical or infeasible for systems which are high-dimensional and exhibit a large disparity in scales. This is because most numerical integrators must use time-steps of the order of the fastest scales which precludes solutions over long time ranges that are of interest for physical and engineering purposes. In the context of atomistic simulations, practically relevant time scales exceed typical integration steps of $\sim 1fs$ by several orders of magnitude [3]. Furthermore, when numerical solutions of transient PDEs are sought, resolution and accuracy requirements give rise to systems with more than 10^9 degrees of freedom [102, 22, 128, 73] where the integration time steps are slaved by fast reaction rates or high oscillation frequencies. This impedes their solution and frequently constitutes computationally infeasible other important tasks such as stability analysis, sensitivity, design and control.

Multiscale dynamical systems exist independently of the availability of mathematical models. Large numbers of time series appear in financial applications, meteorology, remote sensing where the phenomena of interest unfold also over a large range of time scales [139, 80]. A wealth of time series data is also available in experimental physics and engineering which by themselves or in combination with mathematical models can be useful in analyzing underlying phenomena [106, 88, 108] by deriving reduced, predictive descriptions.

Quite frequently the time evolution of all the observables is irrelevant for physical and practical purposes and the analysis is focused on a reduced set of variables or reaction coordinates $\hat{\boldsymbol{y}}_t = \mathcal{P}(\boldsymbol{y}_t)$ obtained by an appropriate mapping $\mathcal{P} : \mathcal{Y} \to \hat{\mathcal{Y}}$. The goal is then to identify a closed, deterministic or stochastic system of equations with respect to $\hat{\boldsymbol{y}}_t$, e.g. :

$$\frac{d\hat{\boldsymbol{y}}_t}{dt} = \hat{\boldsymbol{f}}(\hat{\boldsymbol{y}}_t), \quad \hat{\boldsymbol{y}}_t \in \hat{\mathcal{Y}} \tag{1.3}$$

In the context of equilibrium theormodynamics where enmsemble averages with respect to the invariant distribution of $\hat{\boldsymbol{y}}_t$ are of interest, coarse-graining amounts to free-energy computations [25]. In the nonequilibrium case and when an invariant distribution exists, a general approach for deriving effective dynamics is based on Mori-Zwanzig projections [146, 68, 27, 28, 29, 34]. Other powerful numerical approaches to identify the dynamical behavior with respect to the reduced coordinates include transition path sampling, the transfer operator approach, the nudged elastic band, the string method, Perron cluster analysis and spectral decompositions [39, 42, 43, 49, 40, 105]. Marked efforts in chemical kinetics have led to an array of computational tools such as computational singular perturbation [98, 99], the intrinsic low-dimensional manifold approach [104, 144] and others [120, 113, 90]. Notable successes in overcoming the timescale dilemma have also been achieved in the context of MD simulations [97, 134, 135, 132] (or Hamiltonian systems in general [112, 100, 127]).

In several problems, physical or mathematical arguments have led analysts to identify a few, salient features and their inter-dependencies that macroscopically describe the behavior of very complex systems consisting of a huge number of individuals/agents/components/degrees of freedom. These variables parameterize a low-dimensional, attracting, invariant, "slow" manifold characterizing the long-term process dynamics [71]. Hence the apparent complexity exhibited in the high-dimensionality and the multiscale character of the original model is a pretext of a much simpler, latent structure that, if revealed, could make the aforementioned analysis tasks much more tractable. The emergence of macroscopic, coherent behavior has been the foundation

of coarse-grained dynamic models that have been successful in a wide range of applications. The coarse-grained parameterization and associated model depend on the analysis objectives and particularly on the time scale one wishes to make predictions. Modern approaches with general applicability such as the Equation-free method [92] or Heterogeneous Multiscale Method (HeMM,[47]) are also based on the availability of reduced coordinates and in the case of HeMM of a macroscopic model which is informed and used in conjunction with the microscale model.

Largely independently of the developments in the fields of computational physics and engineering, the problem of deriving, predictive reduced-order models for a large number of time series that potentially exhibit multiple scales has also been addressed in statistics and machine learning communities [140, 53] with applications in network analysis [33], environmetrics [141], sensor network monitoring [142, 117], moving object tracking [4], financial data analysis [5], computer model emulation [103]. Significant advances have been achieved in modeling [131], forecasting [143] and developing online, scalable algorithms [51, 126, 91, 94, 89, 145]. that are frequently based on the discovery of hidden variables that provide insight to the intrinsic structure of streaming data [54, 35, 110, 109, 85].

The present paper proposes a data-driven alternative that is able to automatically coarse-grain high-dimensional systems without the need of preprocessing and availability of physical insight. The data is most commonly obtained by simulations of the most reliable, finest-scale (microscopic) model available. This is used to infer a lower-dimensional description that captures the dynamic evolution of the system at a coarser scale (i.e. a macroscopic model). The majority of available techniques address separately the problems of identifying appropriate reduced coordinates and the effective dynamics in this lower-dimensional representation. It is easily understood that the solution of one affects the other. We propose a general framework where these two problems are simultaneously solved and coarse-grained models are built from the ground up. We propose procedures that concurrently infer the macroscopic dynamics and their mapping the high-dimensional, fine-scale description. As a result no errors or ambiguity are introduced when the fine-scale model needs to be reinitialized in order to obtain additional simulation data. To that end, we advocate a largely unexplored in computational physics perspective based on the Bayesian paradigm which provides a rigorous foundation for learning from data. It is capable of quantifying inferrential uncertainties and, more importantly, uncertainty due to information loss in the coarse-graining process.

We present a Bayesian state-space model where the reduced, coarse-grained dynamics are parametrized by tractable, low-dimensional dynamical models. These can be viewed as experts offering opinions on the evolution of the high-dimensional observables. Each of these modules could represent a single latent regime and would therefore be insufficient by itself to explain the whole system. As is often the case with real experts, their opinions are valid under very specific regimes. We propose therefore a framework for dynamically synthesizing such models in order to obtain an accurate global representation that retains its interpretability and computational tractability.

An important contribution of the paper, particularly in view of enabling simulations of multiscale systems, is online inference algorithms based on Sequential Monte Carlo that scale linearly with the dimensionality of the observables d (Equation (1.1)). These allow the recursive assimilation of data and re-calibration of the coarse-grained dynamics. The Bayesian framework adopted provides probabilistic predictive esti-

mates that can be employed in the context of adaptive time-integration. Rather than determining integration time steps based on traditional accuracy and stability metrics, we propose using measures of the predictive uncertainty in order to decide how long into the future the coarse-grained model can be used. When the uncertainty associated with the predicitve estimates exceeds the analyst's tolerance, the fine-scale dynamics can be consistently re-initialized in order to obtain additional data that sequentially update the coarse-grained model.

In agreement with some recently proposed methodologies [92, 93], the data-driven strategy can seamlessly interact with existing numerical integrators that are well-understood and reliably implemented in several legacy codes. In addition, it is suitable for problems where observational/experimental data must be fused with mathematical descriptions in a rigorous fashion and lead to improved analysis and prediction tools.

The structure of the rest of the paper is as follows. In Section 2 we describe the proposed framework in relation with the state-of-the-art in dimensionality reduction. We provide details of the probabilistic model proposed in the context of Bayesian State-Space models in Section 2.2. Section 2.3 is devoted to inference and learning tasks which involve a locally-optimal Sequential Monte Carlo sampler and an online Expectation-Maximization scheme. The utilization of the coarse-grained dynamics in the context of a Bayesian (adaptive) time-integration scheme is discussed in 2.4 and numerical examples are provided in Section 3.

2. Proposed Approach.

2.1. From static-linear to dynamic-nonlinear dimensionality reduction.

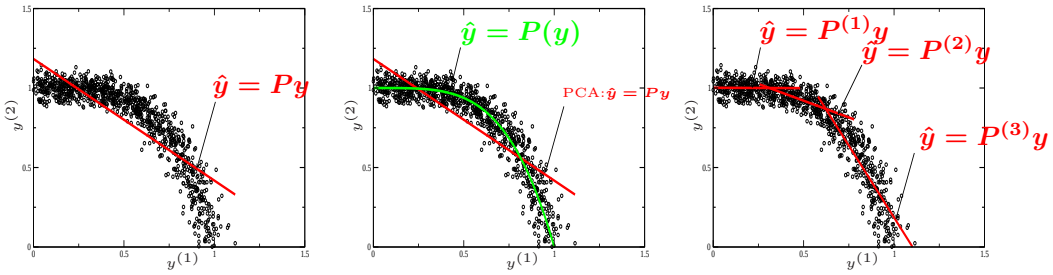
The inherent assumption of all multiscale analysis methods is the existence of a lower-dimensional parameterization of the original system with respect to which the dynamical evolution is more tractable at the scales of interest. In some cases these slow variables can be identified a priori and the problem reduces to finding the necessary closures that will give rise to a consistent dynamical model. In general however one must identify the reduced space $\hat{\mathcal{Y}}$ as well as the dynamics within it.

A prominent role in these efforts has been held by Principal Component Analysis (PCA) -based methods. With small differences and depending on the community other terms such as Proper Orthogonal Decomposition (POD) or Karhunen-Loève expansion (KL), Empirical Orthogonal Functions (EOF) have also been used. PCA finds its roots in the early papers by Pearson [111] and Hotelling [84] and was originally developed as a *static* dimensionality reduction technique. It is based on projections on a reduced basis identified by the leading eigenvectors of the covariance matrix $\boldsymbol{C}$. In the dynamic case and in the absence of closed form expressions for the actual covariance matrix, samples of the process $\boldsymbol{y}_t \in \mathbb{R}^d$ at N distinct time instants t_i are used in order to obtain an estimate of the covariance matrix:

$$\boldsymbol{C} \approx \boldsymbol{C}_N = \frac{1}{N-1} \sum_{i=1}^{N} (\boldsymbol{y}_{t_i} - \boldsymbol{\mu})(\boldsymbol{y}_{t_i} - \boldsymbol{\mu})^T \tag{2.1}$$

where $\boldsymbol{\mu} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{y}_{t_i}$ is the empirical mean. If there is a spectral gap after the first k eigenvalues and $\boldsymbol{V}_K$ is the $d \times K$ matrix whose columns are the K leading normalized eigenvectors of $\boldsymbol{C}_N$ then the reduced-order model is defined with respect to $\hat{\boldsymbol{y}}_t = \boldsymbol{V}_K \boldsymbol{y}_t$. The reduced dynamics can be identified by a Galerkin projection (or a Petrov-Galerkin projection) of the original ODEs in Equation (1.1):

$$\frac{d\hat{\boldsymbol{y}}_t}{dt} = \boldsymbol{V}_K^T \boldsymbol{f}(\boldsymbol{V}_K^T \hat{\boldsymbol{y}}_t) \tag{2.2}$$



(a) PCA (b) Nonlinear PCA (c) Mixture PCA (SLDS)

FIG. 2.1. The phase space is assumed two-dimensional for illustration purposes i.e. $\boldsymbol{y}_t = (y_t^{(1)}, y_t^{(2)})$. Each black circle corresponds to a realization $\boldsymbol{y}_{t_i}$. $\mathcal{P} : \mathcal{Y} \to \hat{\mathcal{Y}}$ is the projection operator from the original high-dimensional space $\mathcal{Y}$ to the reduced-space $\hat{\mathcal{Y}}$.

Hence the reduced space $\hat{\mathcal{Y}}$ is approximated by a hyperplane in $\mathcal{Y}$ and the projection mapping $\mathcal{P}$ linear (Figure 2.1(a)). While it can be readily shown that the projection adopted is optimal in the mean square sense for stationary Gaussian processes, it is generally not so in cases where non-Gaussian processes or other distortion metrics are examined. The application of PCA-based techniques, to high-dimensional, multiscale dynamical systems poses several modeling limitations. Firstly, the reduced space $\hat{\mathcal{Y}}$ might not be sufficiently approximated by a hyperplane of dimension $K << d$. Even though this assumption might hold locally, it is unlikely that this will be a good global approximation. Alternatively, the dynamics of the original process might be adequately approximated on K-dimensional hyperplane but this hyperplane might change in time. Secondly, despite the fact that the projection on the subspace spanned by the leading eigenvectors captures most of the variance of the original process, in cases where this variability is due to the fast modes, there is no guarantee that $\hat{\boldsymbol{y}}_t$ will account for the long-range, slow dynamics which is of primary interest in multiscale systems. Thirdly, the basic assumption in the estimation of the covariance matrix, is that the samples $\boldsymbol{y}_{t_i}$ are drawn from the same distribution, i.e. that the process $\boldsymbol{y}_t$ is stationary. A lot of multiscale problems however involve non-stationary dynamics (e.g. non-equilibrium MD [78, 30]). Hence even if a stationary reduced-order process provides a good, *local*, approximation to $\boldsymbol{y}_t$, it might need to be updated in time. Apart from the aforementioned modeling issues, significant computational difficulties are encountered for high-dimensional systems ($d = dim(\mathcal{Y}) >> 1$) and large datasets ($N >> 1$) as the K leading eigenvectors of large matrices (of dimension proportional to d or N) need to be evaluated. This effort must be repeated, if more samples become available (i.e. N increases) and an update of the reduced-order model is desirable. Recent efforts have concentrated on developing online versions [137] that circumvent this problem.

The obvious extension to the linear projections of PCA is nonlinear dimensionality reduction techniques. These have been the subject of intense research in statistics and machine learning in recent years ([121, 115, 130, 44, 123, 10, 9]) and fairly recently have found their way to computational physics and multiscale dynamical systems (e.g. [32, 96, 107, 59]). They are generally based on calculating eigenvectors of an affinity matrix of a weighted graph. While they circumvent the limiting, linearity assumption of standard PCA, they still assume that the underlying process is stationary (Figure 2.1(b)). Even though the system's dynamics might be appropriately tracked on a

lower-dimensional subspace for a certain time period, this might not be invariant across the whole time range of interest. The identification of the dynamics on the reduced-space $\hat{\mathcal{Y}}$ is not as straightforward as in standard PCA and in most cases, a deterministic or stochastic model is fit directly to the projected data points [31, 50, 58]. More importantly since the inverse mapping $\mathcal{P}^{-1}$ from the manifold $\hat{\mathcal{Y}}$ to $\mathcal{Y}$ is not available analytically, approximations have to be made in order to find pre-images in the data-space [11, 50]. From a computational point of view, the cost of identifying the projection mapping is comparable to standard PCA as an eigenvalue problem on a $N \times N$ matrix has to be solved. Updating those eigenvalues and the nonlinear projection operator in cases where additional data become available implies a significant computational overhead although recent efforts [122] attempt to overcome this limitation.

A common characteristic of the aforementioned techniques is that even though the reduced coordinates are learned from *a finite amount of simulation data*, there is no *quantification of the uncertainty* associated with these inferences. This is a critical component not only in cases where multiple sets of reduced parameters and coarse-grained models are consistent with the data, but also for assessing errors associated with the analysis and prediction estimates. It is one of the main motivations for adopting a *probabilistic approach* in this project. Statistical models can naturally deal with stochastic systems that frequently arise in a lot of applications. Most importantly perhaps, even in cases where the fine-scale model is deterministic (e.g. Equation (1.1)), a stochastic reduced model provides a better approximation that can simultaneously quantify the uncertainty arising from the information loss that takes place during the coarse-graining process [52, 95].

A more general perspective is offered by latent variable models where the observed data (experimental or computationally generated) is augmented by a set of hidden variables [13]. *In the case of high-dimensional, multiscale dynamical systems, the latent model corresponds to a reduced-order process that evolves at scales of practical relevance.* Complex distributions over the observables can be expressed in terms of simpler and tractable joint distributions over the expanded variable space. Furthermore, *structural characteristics* of the original, high-dimensional process $\boldsymbol{y}_t$ can be revealed by interpreting the latent variables as generators of the observables.

In that respect, a general setting is offered by Hidden Markov Models (HMM, [64]) or more generally State-Space Models (SSM) [18, 65, 80]. These assume the existence of an *unobserved (latent)* process $\hat{\boldsymbol{y}}_t \in \mathbb{R}^K$ described by a (stochastic) ODE:

$$\frac{d\hat{\boldsymbol{y}}_t}{dt} = \hat{\boldsymbol{f}}(\hat{\boldsymbol{y}}_t; \boldsymbol{w}_t) \quad \text{(transition equation)} \tag{2.3}$$

which gives rise to the observables $\boldsymbol{y}_t \in \mathbb{R}^d$ as:

$$\boldsymbol{y}_t = \boldsymbol{h}(\hat{\boldsymbol{y}}_t, \boldsymbol{v}_t) \quad \text{(emission equation)} \tag{2.4}$$

where $\boldsymbol{w}_t$ and $\boldsymbol{v}_t$ are unknown stochastic processes (to be inferred from data) and $\hat{\boldsymbol{f}} : \mathbb{R}^K \to \mathbb{R}^K$, $\boldsymbol{h} : \mathbb{R}^K \to \mathbb{R}^d$ are unknown measurable functions. The transition equation defines a prior distribution on the coarse-grained dynamics whereas the emission equation, the mapping that connects the reduced-order representation with the observable dynamics. The object of Bayesian inference is to learn the unobserved (unknown) model parameters from the observed data. Hence the coarse-grained model and its relation to the observable dynamics are inferred from the data.

The form of Equations (2.3) and (2.4) affords general representations. Linear and nonlinear PCA models arise as special cases by appropriate selection of the functions and random processes appearing in the transition and emission equations. Note for example that the transition equation (Equation (2.3)) for $\hat{\boldsymbol{y}}_t$ in the case of the PCA-based models reviewed earlier is given by Equation (2.2) and the *emission equation* (Equation (2.4)) that relates latent and observed processes is linear, deterministic and specified by the matrix of K leading eigenvectors $\boldsymbol{V}_K$.

An extension to HMM is offerered by switching-state models [74, 24, 72, 124] which can be thought of as dynamical mixture models [23, 66]. The latent dynamics consist of a discrete process that takes M values, each corresponding to a distinct dynamical behavior. This can be represented by an M-dimensional vector $\boldsymbol{z}_t$ whose entries are zero except for a single one m which is equal to one and represents the active mode/cluster. Most commonly, the time-evolution of $\boldsymbol{z}_t$ is modeled by a first-order stationary Markov process:

$$\boldsymbol{z}_{t+1} = \boldsymbol{T}\boldsymbol{z}_t \tag{2.5}$$

where $\boldsymbol{T} = [T_{m,n}]$ is the transition matrix and $T_{m,n} = Pr[z_{m,t+1} = 1 \mid z_{n,t} = 1]$. In addition to $\boldsymbol{z}_t$, M processes $\boldsymbol{x}_t^{(m)} \in \mathbb{R}^K,\ m = 1, \ldots, M$ parameterize the reduced-order dynamics (see also discussion in section 2.2). Each is activated when $z_{m,t} = 1$. In the linear version (Switching Linear Dynamic System, SLDS [1]) and conditioned on $z_{m,t} = 1$, the observables $\boldsymbol{y}_t$ arise by a projection from the active $\boldsymbol{x}_t^{(m)}$ as follows:

$$\boldsymbol{y}_t = \boldsymbol{P}^{(m)}\boldsymbol{x}_t^{(m)} + \boldsymbol{v}_t, \quad \boldsymbol{v}_t \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})\ (i.i.d) \tag{2.6}$$

where $\boldsymbol{P}^{(m)}$ are $d \times K$ matrices ($K << d$) and $\boldsymbol{\Sigma}$ is a positive definite $d \times d$ matrix. Such models provide a natural, physical interpretation according to which the behavior of the original process $\boldsymbol{y}_t$ is segmented into M regimes or clusters, the dynamics of which can be low-dimensional and tractable. From a modeling point of view, the idea of utilizing a mixture of simple models provides great flexibility [16, 14, 125, 70] as it can be theorized that given a large enough number of such components, any type of dynamics can be sufficiently approximated. In practice however, a large number might be needed, resulting in complex models containing a large number of parameters.

Such mixture models have gained prominence in recent years in the machine learning community. In [15] for example, a dynamic mixture model was used to analyze a huge number of time series, each corresponding to a word in the English vocabulary as they appear in papers in the journal *Science*. The latent discrete variables represented *topics* and each topic implied a distribution on the space of words. As a result, not only a predictive summary (dimensionality reduction) of the high-dimensional observables was achieved but also an insightful deconstruction of the original time series was made possible. In fact current research in statistics has focused on *infinite* mixture models where the number of components can be automatically inferred from the data ([129, 21, 12, 56, 57]). In the context of computer simulations of high-dimensional systems, such models have been employed by [55, 82, 80, 79, 81] where maximum likelihood techniques were used to learn the model parameters.

In the next sections we present a novel model that generalizes SLDS. Unlike mixture models which assume that $\boldsymbol{y}_t$ is the result of a single reduced-order process at a time, we propose a *partial-membership* model (referred to henceforth as

[1] sometimes referred to as jump-linear or conditional Gaussian models

Partial-Membership Linear Dynamic System, PMLDS) which allows observables to have fractional memberships in multiple clusters. The latent, building blocks are *experts* [86, 87, 77, 75] which, on their own, provide an incomplete, biased prediction but when their "opinions" are appropriately synthesized, they can give rise to a highly accurate aggregate model.

From a modeling perspective such an approach has several appealing properties. The integrated coarse-grained model can be interpretable and low-dimensional even for large, multiscale systems as its expressive ability does not hinge upon the complexity of the individual components but rather is a result of its factorial character ([67]). Intricate dynamical behavior can be captured and decomposed in terms of simple building blocks. It is highly-suited for problems that lack *scale separation* and where the evolution of the system is the result of phenomena at a *cascade of scales*. Each of these scales can be described by a latent process and the resulting coarse-grained model will account not only for the *slow* dynamics but also quantify the predictive uncertainty due to the condensed, fast-varying features.

From an algorithmic point of view we present *parallelizable, online* inference/learning schemes, which can recursively update the estimates produced as more data become available i.e. if the time horizon t of the observables $\boldsymbol{y}_{1:t}$ increases. Unlike some statistical applications where long time series are readily available, in the majority of problems involving computational simulations of high-dimensional, multiscale systems, data is expensive (at least over large time horizons) as they imply calls to the microscopic solvers. The algorithms presented are capable of producing predictive estimates "on the fly" and if additional data is incorporated, they can readily update the model parameters. In addition, such schemes can take advantage of the natural tempering effect of introducing the data sequentially which can further facilitate the solution of the global estimation problem. More importantly perhaps, the updating schemes discussed have *linear complexity* with respect to the dimensionality d of the original process $\boldsymbol{y}_t$.

2.2. Partial-Membership Linear Dynamic System. We present a hierarchical Bayesian framework which promotes sparsity, interpretability and efficiency. The framework described can integrate heterogeneous building blocks and allows for physical insight to be introduced on a case-by-case basis. When dealing with high-dimensional molecular ensembles for example, each of these building blocks might be an (overdamped) Langevin equation with a harmonic potential [17, 80, 83]. It is obvious that such a simplified model would perhaps provide a good approximation under specific, limiting conditions (e.g. at a persistent metastable state) but definitely not across the whole time range of interest. Due to its simple parameterization and computational tractability, it can easily be trained to represent one of the "experts" in the integrated reduced-model. It is known that these models work well under specific regimes but none of them gives an accurate global description. In the framework proposed, they can however be utilized in a way that combines their strengths, but also probabilistically quantifies their limitations.

A transient, nonlinear PDE can be resolved into several linear PDEs whose simpler form and parameterization makes them computational tractable over macroscopic time scales and permits a coarser spatial discretization. Their combination with time-varying characteristics can give rise to an accurate global approximation. Their simplicity and limited range of applicability would preclude their individual use. In the framework proposed however, these simple models would only serve as good local approximants and their inaccurate predictions would be synthesized into an accurate,

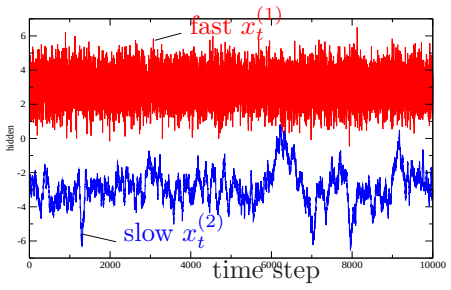


FIG. 2.2. Realizations of two hidden ($M = 2$) one-dimensional ($K = 1$) Ornstein-Uhlenbeck processes were used $d\boldsymbol{x}_t^{(m)} = -b^{(m)}(\boldsymbol{x}_t^{(m)} - \boldsymbol{q}_x^{(m)})dt + (\boldsymbol{\Sigma}^{(m)})^{1/2}d\boldsymbol{W}_t$ with $(b^{(1)}, \boldsymbol{q}_x^{(1)}, \boldsymbol{\Sigma}^{(1)}) = (1., 3., 2.)$ (fast) and $(b^{(2)}, \boldsymbol{q}_x^{(2)}, \boldsymbol{\Sigma}^{(2)}) = (0.01, -3., 0.02)$ (slow)

global model.

2.2.1. Representations with simple building blocks. We describe a probabilistic, dynamic, continuous-time, generative model which relates a sequence the observations $\boldsymbol{y}_t \in \mathbb{R}^d$ at discrete time instants $t = 1, 2, \ldots \tau$ with a number of hidden processes. The proposed model consists of M hidden processes $\boldsymbol{x}_t^{(m)} \in \mathbb{R}^K$, $m = 1, \ldots, M$ ($K << d$) which are assumed to evolve independently of each other and are described by a set of (stochastic) ODEs:

$$\frac{d\boldsymbol{x}_t^{(m)}}{dt} = \boldsymbol{g}_m(\boldsymbol{x}_t^{(m)}; \boldsymbol{\theta}_x^{(m)}), \quad m = 1, \ldots, M \tag{2.7}$$

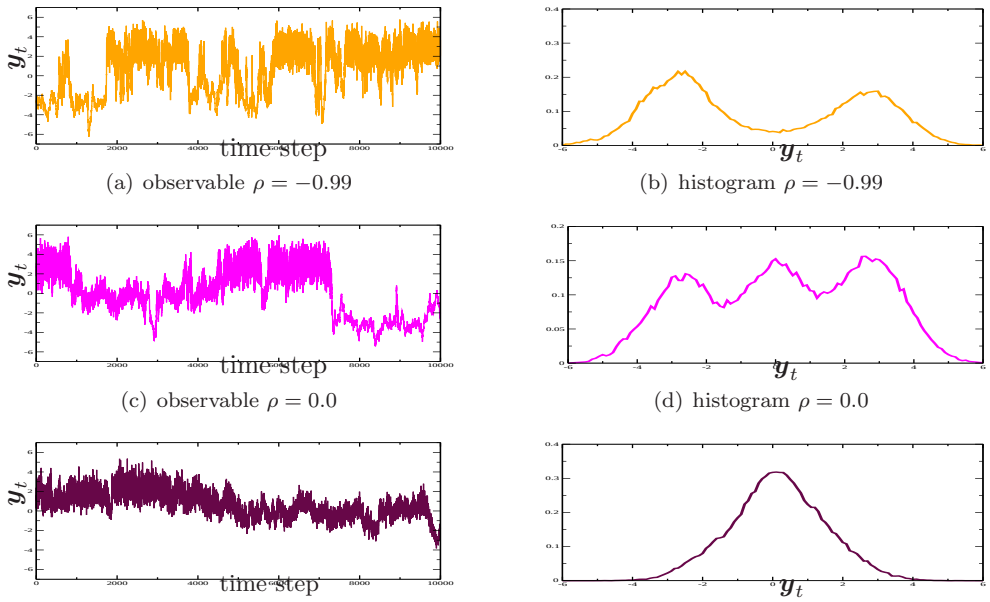
This equation essentially implies a *prior distribution* on the space of hidden processes parameterized by a set of (unknown a priori) parameters $\boldsymbol{\theta}_x^{(m)}$. It should be noted that while the proposed framework allows for any type of process in Equation (2.7), it is desirable that these are simple, in the sense that the parameters $\boldsymbol{\theta}_x^{(m)}$ are low-dimensional and can be learned swiftly and efficiently. We list some desirable properties of the prior models [125]:

- Stationarity: Unless specific prior information is available, it would be unreasonable to impose a time bias on the evolution of any of the reduced dynamics processes. Hence it is important that the models adopted are a priori stationary. Note that the posterior distributions might still exhibit non-stationarity.
- Correlation Decay: It is easily understood that for any m and t_1, t_2, the correlation $\boldsymbol{x}_{t_1}^{(m)}$ and $\boldsymbol{x}_{t_2}^{(m)}$ should decay monotonically as $\mid t_2 - t_1 \mid$ goes to $+\infty$. This precludes models that do not explicitly account for the time evolution of the latent processes and assume that hidden states are not time-dependent (e.g. static PCA models).
- Other: Although this is not necessary, we adopt a continuous time model in the sense of [136] with an analytically available transition density which allows inference to be carried out seamlessly even in cases where the observables are obtained at non-equidistant times. As a result the proposed framework can adapt to the granularity of the observables and also provide exact probabilistic predictions at any time resolution.

Although more complex models can be adopted we assume here that independent, isotropic Ornstein-Uhlenbeck (OU) processes are used to model the hidden dynamics $\boldsymbol{x}_t^{(m)}$. The OU processes used comply with the aforementioned desiderata. In particular, the following parameterization is employed:

$$d\boldsymbol{x}_t^{(m)} = -b_x^{(m)}(\boldsymbol{x}_t - \boldsymbol{q}_x^{(m)})dt + (\boldsymbol{S}_x^{(m)})^{1/2}d\boldsymbol{W}_t^{(m)} \tag{2.8}$$

where $\boldsymbol{W}_t^{(m)}$ are Wiener processes (independent for each m), $b_x^{(m)} > 0$, $\boldsymbol{q}_x^{(m)} \in \mathbb{R}^K$ and $\boldsymbol{S}_x^{(m)}$ are positive definite matrices of dimension $K \times K$. The aforementioned



(a) observable $\rho = -0.99$ (b) histogram $\rho = -0.99$

(c) observable $\rho = 0.0$ (d) histogram $\rho = 0.0$

(e) observable $\rho = 0.99$ (f) histogram $\rho = 0.99$

FIG. 2.3. The logistic normal distribution was used to model the weights associated with each of the hidden processes depicted in Figure 2.2. In particular, an isotropic Ornstein-Uhlenbeck process $d\mathbf{z}_t = -b_z(\mathbf{z}_t - \boldsymbol{q}_z)dt + \boldsymbol{\Sigma}_z^{1/2} d\boldsymbol{W}_t$ with $b_z = 0.001$, $\boldsymbol{q}_z = [0,\ 0]^T$ and $\boldsymbol{\Sigma}_z$. Graphs depict the resulting observable time history (left column) and its histogram (right column) arising from the unobserved processes in Figure 2.2 and for three values of ρ. It is noted that time histories exhibit fast and slow scales of the processes in Figure 2.2. Furthermore, by changing a single parameter (i.e. ρ) one can obtain two, three or a single *metastable* state (right column - peaks of the histogram).

model has a Gaussian invariant (stationary) distribution $\mathcal{N}(\boldsymbol{q}_x^{(m)}, \frac{1}{2b_x^{(m)}}\boldsymbol{S}_x^{(m)})$. The transition density denoted by $p(\boldsymbol{x}_t^{(m)} \mid \boldsymbol{x}_{t-1}^{(m)})$ for time separation δt is also a Gaussian $\mathcal{N}(\boldsymbol{\mu}_{t,\delta t}, \boldsymbol{S}_{\delta t})$ where:

$$\begin{aligned} \boldsymbol{\mu}_{t,\delta t} &= \boldsymbol{x}_{t-1}^{(m)} - (1 - e^{-b_x^{(m)}\delta t})(\boldsymbol{x}_{t-1}^{(m)} - \boldsymbol{q}_x^{(m)}) \\ \boldsymbol{S}_{\delta t} &= \frac{1-e^{-2b_x^{(m)}\delta t}}{2b_x^{(m)}}\boldsymbol{S}_x^{(m)} \end{aligned} \tag{2.9}$$

It is not expected that simple processes on their own will provide good approximations to the essential dynamics exhibited in the data $\boldsymbol{y}_t$. In order to combine the dynamics implied by the M processes in Equation (2.7), we consider an $M-$dimensional process $\mathbf{z}_t$ such that $\sum_{m=1}^M z_{m,t} = 1$ and $z_{m,t} > 0,\ \forall t$ and define an appropriate prior. The coefficients $z_{m,t}$ express the *weight* or *fractional membership* to each process/expert $\boldsymbol{x}_t^{(m)}$ at time t [76]. We use the *logistic-normal* model [15] as a prior for $\mathbf{z}_t$. It is based on a Gaussian process, $\hat{\boldsymbol{z}}_t$ whose dynamics are also prescribed by an isotropic OU process:

$$d\hat{\boldsymbol{z}}_t = -b_z(\hat{\boldsymbol{z}}_t - \boldsymbol{q}_z)dt + \boldsymbol{S}_z^{1/2} d\boldsymbol{W}_t \tag{2.10}$$

and the transformation:

$$z_{m,t} = \frac{e^{\hat{z}_{m,t}} + 1/M}{\sum_{m=1}^M e^{\hat{z}_{m,t}} + 1},\ \forall m, t \tag{2.11}$$

The invariant and transition densities of $\hat{\boldsymbol{z}}_t$ are obviously identical to the ones for $\boldsymbol{x}_t^{(m)}$ with appropriate substitution of the parameters. The hidden processes $\{\boldsymbol{x}_t^{(m)}\}_{m=1}^M$ and associated weights $\mathbf{z}_t$ give rise to the observables $\boldsymbol{y}_t$ as follows (compare with Equation (2.6)):

$$\boldsymbol{y}_t = \sum_{m=1}^{M} z_{m,t}\ \boldsymbol{P}^{(m)}\boldsymbol{x}_t^{(m)} + \boldsymbol{v}_t, \quad \boldsymbol{v}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma})\ (i.i.d) \tag{2.12}$$

where $\boldsymbol{P}^{(m)}$ are $d \times K$ matrices ($K << d$) and $\boldsymbol{\Sigma}$ is a positive definite $d \times d$ matrix. The aforementioned equation implies a series of linear projections on hyperplanes of dimension K. The dynamics along those hyperplanes are dictated by a priori independent process $\boldsymbol{x}_t^{(m)}$. It is noted however the reduced dynamics are simultaneously described by *all* the hidden processes (Figure 2.4). This is in contrast to PCA methods where a single such projection is considered and mixture PCA models where even though several hidden processes are used, at each time instant it is assumed that a single one is active. Due to the factorial nature of the proposed model, multiple dynamic regimes can be captured by appropriately combining a few latent states. While mixture models (Figure 2.1(c)) provide a flexible framework, the number of hidden states might be impractically large. As it is pointed out in [67], in order to encode for example a time sequence with $30bits$ of information one would need $k = 2^{30}$ distinct states. It is noted that even though *linear* projections are implied by Equation (2.12) and Gaussian noise $\boldsymbol{v}_t$ is used, the resulting model for $\boldsymbol{y}_t$ is *nonlinear* and *non-Gaussian* as it involves the factorial combination of M processes $\{\boldsymbol{x}_t^{(m)}\}_{m=1}^M$ with $\mathbf{z}_t$ which are *a posteriori* non-Gaussian.

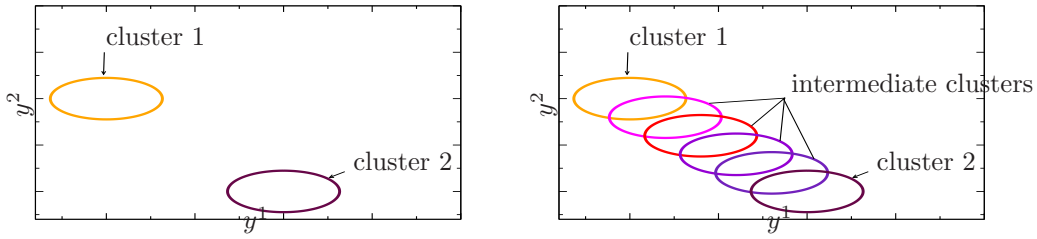
The parameters $z_{m,t}$ express the relative importance of the various reduced models or equivalently the degree to which each data point $\boldsymbol{y}_t$ is associated with each of the M reduced dynamics $\boldsymbol{x}_t^{(m)}$. It is important to note that the proposed model allows for time varying weights $\boldsymbol{z}_{m,t}$ and can therefore account for the possibility of switching between different regimes of dynamics. Figure 2.3 depicts a simple example ($d = 1$) which illustrates the flexibility of the proposed approach.

The unknown parameters of the coarse-grained model consist of:

- *dynamic* variables denoted for notational economy by $\boldsymbol{\Theta}_t$ (i.e. $\{\boldsymbol{x}_t^{(m)}\}_{m=1}^M$, $\mathbf{z}_t$, for $t = 1, 2, \ldots$).
- *static* variables denoted by $\boldsymbol{\Theta}$ (i.e. $\boldsymbol{\theta}_x^{(m)} = (b_x^{(m)}, \boldsymbol{q}_x^{(m)}, \boldsymbol{S}_x^{(m)})$ in Equation (2.8), $\boldsymbol{\theta}_z = (b_z, \boldsymbol{q}_z, \boldsymbol{S}_z)$ in Equation (2.10) and $\{\boldsymbol{P}^{(m)}\}_{m=1}^M, \boldsymbol{\Sigma}$ in Equation (2.12)).

2.3. Inference and learning. Inference in the probabilistic graphical model described involves determining the probability distributions associated with the unobserved (hidden) static $\boldsymbol{\Theta}$ and dynamic parameters $\boldsymbol{\Theta}_t$ of the model. In the Bayesian setting adopted this is the *posterior* distribution of the unknown parameters of the coarse-grained model. Given the observations (computational or experimental) of the original, high-dimensional process $\boldsymbol{y}_{1:\tau} = \{\boldsymbol{y}_t\}_{t=1}^{\tau}$, we denote the posterior by $\pi(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{1:\tau})$:

$$\pi(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{1:\tau}) = p(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{1:\tau} \mid \boldsymbol{y}_{1:\tau}) = \frac{\overbrace{p(\boldsymbol{y}_{1:\tau} \mid \boldsymbol{\Theta}, \boldsymbol{\Theta}_{1:\tau})}^{likelihood}\ \overbrace{p(\boldsymbol{\Theta}_t, \boldsymbol{\Theta})}^{prior}}{p(\boldsymbol{y}_{1:\tau})} \tag{2.13}$$



(a) SLDS with two clusters (b) proposed Partial-Membership LDS

FIG. 2.4. *Comparison of SLDS (a) with two mixture components and the proposed model partial-membership model (b)*

The normalization constant $p(\boldsymbol{y}_{1:\tau})$ is not of interest when sampling for $\boldsymbol{\Theta}$ or $\boldsymbol{\Theta}_{1:\tau}$ but can be quite useful for model validation purposes.

A telescopic decomposition holds for the likelihood which according to Equation (2.12) is given by:

$$p(\boldsymbol{y}_{1:\tau} \mid \boldsymbol{\Theta}, \boldsymbol{\Theta}_{1:\tau}) = \prod_{t=1}^{\tau} p(\boldsymbol{y}_t \mid \boldsymbol{\Theta}, \boldsymbol{\Theta}_t) \tag{2.14}$$

where the densities in the product are described in Equation (2.17). Equation (2.12) defines the likelihood $p(\boldsymbol{y}_t \mid \boldsymbol{\Theta}, \boldsymbol{\Theta}_t)$ which is basically the *weighted product* of the likelihoods under each of the hidden processes/experts $\boldsymbol{x}_t^{(m)}$:

$$p(\boldsymbol{y}_t \mid \boldsymbol{\Theta}, \boldsymbol{\Theta}_t) = \frac{1}{c(\boldsymbol{\Theta}, \boldsymbol{\Theta}_t)} \prod_{m=1}^{M} p_m^{z_{m,t}}(\boldsymbol{y}_t \mid \boldsymbol{\Theta}, \boldsymbol{\Theta}_t) \tag{2.15}$$

where the normalizing constant $c(\boldsymbol{\Theta}, \boldsymbol{\Theta}_t)$ ensures that the density intergrates to one with respect to $\boldsymbol{y}_t$. According to Equation (2.12):

$$p_m(\boldsymbol{y}_t \mid \boldsymbol{\Theta}, \boldsymbol{\Theta}_t) \propto \frac{1}{\mid \boldsymbol{\Sigma} \mid^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{y}_t - \boldsymbol{P}^{(m)} \boldsymbol{x}_t^{(m)})^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_t - \boldsymbol{P}^{(m)} \boldsymbol{x}_t^{(m)}) \right\} \tag{2.16}$$

The likelihood can be written in a more compact form as:

$$p(\boldsymbol{y}_t \mid \boldsymbol{\Theta}, \boldsymbol{\Theta}_t) \propto \frac{1}{\mid \boldsymbol{\Sigma} \mid^{1/2}} \exp\left\{ -\frac{1}{2} (\boldsymbol{y}_t - \boldsymbol{W}_t \boldsymbol{X}_t)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_t - \boldsymbol{W}_t \boldsymbol{X}_t) \right\} \tag{2.17}$$

where:

$$\underbrace{\boldsymbol{X}_t^T}_{MK \times 1} = \left[(\boldsymbol{x}_t^{(1)})^T, (\boldsymbol{x}_t^{(2)})^T, \ldots (\boldsymbol{x}_t^{(M)})^T \right]^T \tag{2.18}$$

and:

$$\underbrace{\boldsymbol{W}_t}_{d \times MK} = \left[\begin{array}{cccccccc} z_{1,t} & \boldsymbol{P}^{(1)} & z_{2,t} & \boldsymbol{P}^{(2)} & \ldots & z_{M,t} & \boldsymbol{P}^{(M)} \end{array} \right] \tag{2.19}$$

The first-order Markovian processes adopted for the prior modeling of the dynamic parameters $\boldsymbol{\Theta}_t$ (Equations (2.8), (2.10), 2.9)) imply that:

$$p(\boldsymbol{\Theta}_{1:\tau}, \boldsymbol{\Theta}) = p(\boldsymbol{\Theta}) \prod_{t=1}^{\tau} p(\boldsymbol{\Theta}_t \mid \boldsymbol{\Theta}_{t-1}, \boldsymbol{\Theta}) \tag{2.20}$$

where $p(\boldsymbol{\Theta}_1 \mid \boldsymbol{\Theta}_0, \boldsymbol{\Theta}) = p(\boldsymbol{\Theta}_1 \mid \boldsymbol{\Theta}) = \nu_0(\boldsymbol{\Theta}_1 \mid \boldsymbol{\Theta})$ is the prior on the initial condition which in this work is taken to be the stationary distribution of the underlying OU processes (see discussion in Section 2.2.1) and denoted for notational economy by $\nu_0(. \mid \boldsymbol{\Theta})$.

The posterior encapsulates uncertainties arising from the potentially stochastic nature of the original process $\boldsymbol{y}_t$ as well as due to the fact that a finite number of observations were used. The difficulty of the problem is that both the dynamic ($\boldsymbol{\Theta}_{1:\tau}$) and the static parameters ($\boldsymbol{\Theta}$) are unknown. We adopt a hybrid strategy whereby we sample from the full posterior for the dynamic parameters $\boldsymbol{\Theta}_t$ and provide point estimates for the static parameters $\boldsymbol{\Theta}$. If uniform priors are used for $\boldsymbol{\Theta}$ then the procedure proposed reduces to a maximum likelihood estimation. Non-uniform priors have a regularization effect which can promote the identification of particular features.

While the hybrid strategy proposed is common practice in pertinent models ([64]), in the current framework it is also necessitated by the difficulty in sampling in the high-dimensional state space of $\boldsymbol{\Theta}$ (note that the projection matrices $\boldsymbol{P}^{(m)}$ in particular are of the dimension of the observables d and $d >> 1$) as well as the need for scalability in the context of high-dimensional systems. The static parameters $\boldsymbol{\Theta}$ are estimated by maximizing the *log-posterior*.

$$L(\boldsymbol{\Theta}) = \log \pi(\boldsymbol{\Theta} \mid \boldsymbol{y}_{1:\tau}) = \log \int \underbrace{\pi(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{1:\tau} \mid \boldsymbol{y}_{1:\tau})}_{posterior\ \ Equation(2.13)} \ d\boldsymbol{\Theta}_{1:\tau} \tag{2.21}$$

Maximization of $L(\boldsymbol{\Theta})$ is more complex than a standard optimization task as it involves integration over the unobserved dynamic variables $\boldsymbol{\Theta}_{1:\tau}$. While maximization can be accelerated by using gradient-based techniques (e.g. gradient ascent), the dimensionality of $\boldsymbol{\Theta}$ makes such an approach impractical as it can be extremely difficult to scale the parameter increments. We propose therefore adopting an Expectation-Maximization framework (EM) which is an iterative, robust scheme that is guaranteed to increase the log-posterior at each iteration ([41, 64]). It is based on constructing a series of increasing lower bounds of the log-posterior using auxiliary distributions $q(\boldsymbol{\Theta}_{1:\tau})$:

$$\begin{aligned} L(\boldsymbol{\Theta}) = \log \pi(\boldsymbol{\Theta} \mid \boldsymbol{y}_{1:\tau}) \quad &= \log \textstyle\int \pi(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{1:\tau} \mid \boldsymbol{y}_{1:\tau}) \ d\boldsymbol{\Theta}_{1:\tau} \\ &= \log \textstyle\int q(\boldsymbol{\Theta}_{1:\tau}) \frac{\pi(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{1:\tau} \mid \boldsymbol{y}_{1:\tau})}{q(\boldsymbol{\Theta}_{1:\tau})} \ d\boldsymbol{\Theta}_{1:\tau} \\ &\geq \textstyle\int q(\boldsymbol{\Theta}_{1:\tau}) \log \frac{\pi(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{1:\tau} \mid \boldsymbol{y}_{1:\tau})}{q(\boldsymbol{\Theta}_{1:\tau})} \ d\boldsymbol{\Theta}_{1:\tau} \quad \text{(Jensen's inequality)} \\ &= F(q, \boldsymbol{\Theta}) \end{aligned} \tag{2.22}$$

It is obvious that this inequality becomes an equality when in place of the auxiliary distribution $q(\boldsymbol{\Theta}_{1:\tau})$ the posterior $\pi(\boldsymbol{\Theta}_{1:\tau} \mid \boldsymbol{\Theta}, \boldsymbol{y}_{1:\tau})$ is selected. Given an estimate $\boldsymbol{\Theta}^{(s)}$ at step s, this suggests iterating between an Expectation step (E-step) whereby we average with respect to $q^{(s)}(\boldsymbol{\Theta}_{1:\tau}) = \pi(\boldsymbol{\Theta}_{1:\tau} \mid \boldsymbol{\Theta}^{(s)}, \boldsymbol{y}_{1:\tau})$ to evaluate the lower bound:

$$\begin{aligned} \textbf{E-step:} \quad F^{(s)}(q^{(s)}, \boldsymbol{\Theta}) \quad &= \textstyle\int q^{(s)}(\boldsymbol{\Theta}_{1:\tau}) \log \pi(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{1:\tau} \mid \boldsymbol{y}_{1:\tau}) \ d\boldsymbol{\Theta}_{1:\tau} \\ &\quad - \textstyle\int q^{(s)}(\boldsymbol{\Theta}_{1:\tau}) \log q^{(s)}(\boldsymbol{\Theta}_{1:\tau}) \ d\boldsymbol{\Theta}_{1:\tau} \end{aligned} \tag{2.23}$$

and a Maximization step (M-step) with respect to $F^{(s)}(q^{(s)}, \boldsymbol{\Theta})$ (and in particular the

first part in Equation (2.23) since the second does no not depend on $\boldsymbol{\Theta}$):

$$\begin{aligned} \textbf{M-step:} \quad \boldsymbol{\Theta}^{(s+1)} &= arg\max_{\boldsymbol{\Theta}} F^{(s)}(q^{(s)}, \boldsymbol{\Theta}) \\ &= arg\max_{\boldsymbol{\Theta}} E_{q^{(s)}(\boldsymbol{\Theta}_{1:\tau})}\left[\log \pi(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{1:\tau} \mid \boldsymbol{y}_{1:\tau})\right] \\ &= arg\max_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}^{(s)}, \boldsymbol{\Theta}) \end{aligned} \tag{2.24}$$

As the optimal auxiliary distributions $q^{(s)}(\boldsymbol{\Theta}_{1:\tau}) = \pi(\boldsymbol{\Theta}_{1:\tau} \mid \boldsymbol{\Theta}^{(s)}, \boldsymbol{y}_{1:\tau})$ are intractable, we propose employing a Sequential Monte Carlo (SMC or particle filter, [45, 37]) scheme for estimating the expectations in the M-Step, i.e. Equation (2.24). SMC samplers provide a *parallelizable* framework for non-linear, non-Gaussian filtering problems whereby the target distribution $q^{(s)}(\boldsymbol{\Theta}_{1:\tau}) = \pi(\boldsymbol{\Theta}_{1:\tau} \mid \boldsymbol{\Theta}^{(s)}, \boldsymbol{y}_{1:\tau})$ is represented with a population of N particles $\boldsymbol{\Theta}_{1:\tau}^{(s,i)}$ and weights $W^{(s,i)}$ such that the expectation in Equation (2.24) can be approximated by:

$$E_{q^{(s)}(\boldsymbol{\Theta}_{1:\tau})}\left[\log \pi(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{1:\tau} \mid \boldsymbol{y}_{1:\tau})\right] \approx \sum_{i=1}^{N} W^{(s,i)} \ \log \pi(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{1:\tau}^{(s,i)}) \mid \boldsymbol{y}_{1:\tau}) \tag{2.25}$$

In section 2.3.1 we discuss a particle filter that takes advantage of the particular structure of the posterior and employs the locally optimal importance sampling distribution. Nevertheless, SMC samplers involve sequential importance sampling, and their performance decays with increasing τ as the dimension of the state space $\boldsymbol{\Theta}_{1:\tau}$ increases even when resampling and rejuvenation mechanisms are employed ([7]). Recent efforts based on exponential forgetting have shown that the accuracy of the approximation can be improved (while keeping the number of particles N fixed) by employing *smoothing* ([69]) over a fixed-lag in the past ([20]).

In this paper we make use of an *approximate* but highly efficient alternative proposed in [6, 7, 8]. This is based on the so-called split-data likelihood (SDL) first discussed in [116], which consists of splitting the observations into blocks (overlapping or non-overlapping) of length $L < \tau$ and using the pseudo-likelihood which arises by assuming that these blocks are independent. It is shown in [7] that this leads to an alternative Kullback-Leibler divergence contrast function and under some regularity conditions that the set of parameters optimizing this contrast function includes the true parameter. Because the size of the blocks is fixed, the degeneracy of particle filters can be averted and the quality of the approximations can be further improved by applying a backward smoothing step over each block ([69]). Let k denote the index of the block of length L considered and $\bar{\boldsymbol{y}}_k = \boldsymbol{y}_{(k-1)L+1:kL}$ and $\bar{\boldsymbol{\Theta}}_k = \boldsymbol{\Theta}_{(k-1)L+1:kL}$. If $\tau = r\ L$. The likelihood is approximated by:

$$p(\boldsymbol{y}_{1:\tau} \mid \boldsymbol{\Theta}, \boldsymbol{\Theta}_{1:\tau}) \approx \prod_{k=1}^{r} p(\bar{\boldsymbol{y}}_k \mid \boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}}_k) \tag{2.26}$$

When $\boldsymbol{\Theta}_t$ has reached a stationary regime with invariant density $\nu_0(. \mid \boldsymbol{\Theta})$, then for any k, $(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}}_k, \bar{\boldsymbol{y}}_k)$ are identically distributed according to:

$$\begin{aligned} \bar{p}(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}}_k, \bar{\boldsymbol{y}}_k) = \quad & \pi(\boldsymbol{\Theta})\nu_0(\boldsymbol{\Theta}_{(k-1)L} \mid \boldsymbol{\Theta}) p(\boldsymbol{y}_{(k-1)L} \mid \boldsymbol{\Theta}_{(k-1)L}, \boldsymbol{\Theta}) \\ & \prod_{t=(k-1)L+1}^{kL-1} p(\boldsymbol{\Theta}_t \mid \boldsymbol{\Theta}_{t-1}, \boldsymbol{\Theta}) p(\boldsymbol{y}_t \mid \boldsymbol{\Theta}_t, \boldsymbol{\Theta}) \end{aligned} \tag{2.27}$$

In a batch EM algorithm using the split-data likelihood and the k^{th} block of data, the M-step would involve maximization with respect to $\boldsymbol{\Theta}$ of (see also Equation

(2.24)):

$$\bar{Q}(\boldsymbol{\Theta}^{(k-1)}, \boldsymbol{\Theta}) = \int \bar{p}(\bar{\boldsymbol{\Theta}}_k \mid \boldsymbol{\Theta}^{(k-1)}, \bar{\boldsymbol{y}}_k) \log \bar{p}(\boldsymbol{\Theta}, \bar{\boldsymbol{\Theta}}_k, \bar{\boldsymbol{y}}_k)\ d\bar{\boldsymbol{\Theta}}_k \tag{2.28}$$

We utilize an *online* EM algorithm where the iteration numbers s coincide with the block index k (i.e. $s \equiv k$) which effectively implies that the estimates for $\boldsymbol{\Theta}$ are updated every time a new data block is considered. The expectation step (E-step) is replaced by a stochastic approximation ([119, 19]) while the M-step is left unchanged. In particular, at iteration k ($\equiv s$):

$$\begin{aligned} \textbf{online E-step} \quad \bar{Q}(\boldsymbol{\Theta}^{(1:k-1)}, \boldsymbol{\Theta}) \quad &= (1-\gamma_k)\bar{Q}(\boldsymbol{\Theta}^{(1:k-2)}, \boldsymbol{\Theta}) \\ &+\gamma_k \int \bar{p}(\bar{\boldsymbol{\Theta}}_k \mid \boldsymbol{\Theta}^{(k-1)}, \bar{\boldsymbol{y}}_k) \log \bar{p}(\bar{\boldsymbol{\Theta}}_k, \bar{\boldsymbol{y}}_k)\ d\bar{\boldsymbol{\Theta}}_k \end{aligned} \tag{2.29}$$

and update the value of the parameters $\boldsymbol{\Theta}$ as:

$$\textbf{online M-step} \quad \boldsymbol{\Theta}^{(k)} = arg\max_{\boldsymbol{\Theta}} \bar{Q}(\boldsymbol{\Theta}^{(1:k-1)}, \boldsymbol{\Theta}) \tag{2.30}$$

The algorithm relies on a non-increasing sequence of positive stepsizes $\{\gamma_k\}_{k\geq 0}$ such that $\sum_k \gamma_k = +\infty$ and $\sum_k \gamma_k^2 < +\infty$. In this work we adopted $\gamma_k = \frac{1}{k^a}$ with $a = 0.51$. Naturally the integrals above over the hidden dynamic variables $\bar{\boldsymbol{\Theta}}_k$ are estimated using SMC-based, particulate approximations of $\bar{p}(\bar{\boldsymbol{\Theta}}_k \mid \boldsymbol{\Theta}^{(k-1)}\bar{\boldsymbol{y}}_k)$. For small L the convergence will in general be slow as the split-block likelihood assumption will be further from the truth. For larger L, convergence is faster but the performance of the filter decays. For that purpose we also employed a backward smoothing filter over each block using the algorithm described in [69]. The computational cost of the smoothing algorithm is $O(N^2 L)$.

In practice, and in particular for the exponential distributions utilized in the proposed model (e.g. Equation (2.17)), the EM iterations reduce to calculating a set of (multivariate) sufficient statistics $\boldsymbol{\Phi}$. In particular, instead of the log-posterior lower bound $\bar{Q}(\boldsymbol{\Theta}^{(1:k-1)}, \boldsymbol{\Theta})$ in Equation (2.29) we update the sufficient statistics as follows:

$$\begin{aligned} \boldsymbol{\Phi}^{(k)} \quad &= (1-\gamma_k)\boldsymbol{\Phi}^{(k-1)} \\ &+\gamma_k \int \bar{p}(\bar{\boldsymbol{\Theta}}_k \mid \boldsymbol{\Theta}^{(k-1)}, \bar{\boldsymbol{y}}_k)\boldsymbol{\phi}(\bar{\boldsymbol{\Theta}}_k)\ d\bar{\boldsymbol{\Theta}}_k \end{aligned} \tag{2.31}$$

where $\int \bar{p}(\bar{\boldsymbol{\Theta}}_k \mid \boldsymbol{\Theta}^{(k-1)}, \bar{\boldsymbol{y}}_k)\boldsymbol{\phi}(\bar{\boldsymbol{\Theta}}_k)\ d\bar{\boldsymbol{\Theta}}_k$ denotes the set of sufficient statistics associated with the block of data $\bar{\boldsymbol{y}} = \boldsymbol{y}_{(k-1)L+1:kL}$. Specific details are provided in the Appendix. It is finally noted, that learning tasks in the context of the probabilistic model proposed, should also involve identifying the correct *structure* (e.g. the number of different experts M). While this problem poses some very challenging issues which are currently the topic of active research in various contexts (e.g. nonparametric methods), this paper is exclusively concerned with parameter learning. In section 3, we discuss Bayesian validation techniques for assessing quantitatively the correct model structure which are computationally feasible due to the efficiency of the proposed algorithms.

2.3.1. Locally optimal Sequential Monte Carlo samplers. In this section we discuss Monte Carlo approximations of the expectations appearing in Equation (2.31) with respect to the density $\bar{p}(\bar{\boldsymbol{\Theta}}_k \mid \boldsymbol{\Theta}, \boldsymbol{y}_{1:L}) = p(\boldsymbol{\Theta}_{(k-1)L+1:kL} \mid \boldsymbol{\Theta}, \boldsymbol{y}_{(k-1)L+1:kL})$. Note that in order to simplify the notation we consider an arbitrary block of length L

(e.g. $k=1$)and do not explicitly indicate the iteration number of the EM algorithm. Hence the target density is:

$$p(\boldsymbol{\Theta}_{1:L} \mid \boldsymbol{\Theta}, \boldsymbol{y}_{1:L}) = \frac{1}{p(\boldsymbol{y}_{1:L} \mid \boldsymbol{\Theta})} \nu_0(\boldsymbol{\Theta}_1 \mid \boldsymbol{\Theta}) p(\boldsymbol{y}_1 \mid \boldsymbol{\Theta}_1, \boldsymbol{\Theta}) \prod_{t=2}^{L} p(\boldsymbol{\Theta}_t \mid \boldsymbol{\Theta}_{t-1}, \boldsymbol{\Theta}) p(\boldsymbol{y}_t \mid \boldsymbol{\Theta}_t, \boldsymbol{\Theta}) \tag{2.32}$$

where the dynamic variables are $\boldsymbol{\Theta}_t = (\boldsymbol{X}_t, \boldsymbol{z}_t)$ (Equation (2.18)). Based on earlier discussions, the evolution dynamics of these variables are independent i.e.:

$$\nu_0(\boldsymbol{\Theta}_1 \mid \boldsymbol{\Theta}) = \nu_0(\boldsymbol{X}_1 \mid \boldsymbol{\Theta}) v_0(\boldsymbol{z}_1 \mid \boldsymbol{\Theta}) \tag{2.33}$$

and:

$$p(\boldsymbol{\Theta}_t \mid \boldsymbol{\Theta}_{t-1}, \boldsymbol{\Theta}) = p(\boldsymbol{X}_t \mid \boldsymbol{X}_{t-1}, \boldsymbol{\Theta}) p(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}, \boldsymbol{\Theta}) \tag{2.34}$$

Since there is a deterministic relation between $\bar{\boldsymbol{z}}_t$ and $\boldsymbol{z}_t$ (Equation (2.11)) we use them interchangeably. In particular we use $\hat{\boldsymbol{z}}_t$ in the evolution equations since the initial and transition densities are Gaussian (Equation (2.10)) and $\boldsymbol{z}_t$ in the likelihood densities as the expressions simplify in Equation (2.12)). The initial and transition densities for $\boldsymbol{X}_t$ are also Gaussian. Given that $\boldsymbol{x}_t^{(m)}$ are a priori independent, we have that:

$$\begin{array}{ll} p(\boldsymbol{X}_t \mid \boldsymbol{X}_{t-1}, \boldsymbol{\Theta}) & = \prod_{m=1}^{M} p(\boldsymbol{x}_t^{(m)} \mid \boldsymbol{x}_{t-1}^{(m)}, \boldsymbol{\Theta}) \\ & = \mathcal{N}(\boldsymbol{X}_t \mid \boldsymbol{\mu}_t, \boldsymbol{S}_X) \end{array} \tag{2.35}$$

where the mean $\boldsymbol{\mu}_t = \boldsymbol{\mu}_t(\boldsymbol{X}_{t-1})$ is given by Equation (2.9) and $\boldsymbol{S}_X = \mathrm{diag}(\boldsymbol{S}_{x,1}, \ldots, \boldsymbol{S}_{x,M})$ (from Equation (2.9) as well).

SMC samplers operate on a sequence of target densities $p(\boldsymbol{\Theta}_{1:t} \mid \boldsymbol{y}_{1:t}, \boldsymbol{\Theta})$ which, for any t, are approximated by a set of n random samples (or *particles*) $\{\boldsymbol{\Theta}_{1:t}^{(i)}\}_{i=1}^{n}$. These are propagated forward in time using a combination of *importance sampling*, *resampling* and MCMC-based *rejuvenation* mechanisms [38, 37, 138, 133]. Each of these particles is associated with an *importance weight* $W^{(i)}$ ($\sum_{i=1}^{n} W^{(i)} = 1$) which is updated sequentially along with the particle locations in order to provide a particulate approximation:

$$p(\boldsymbol{\Theta}_{1:t} \mid \boldsymbol{y}_{1:t}, \boldsymbol{\Theta}) \approx \sum_{i=1}^{n} W^{(i)} \, \delta_{\boldsymbol{\Theta}_{1:t}^{(i)}}(\boldsymbol{\Theta}_{1:t}) \tag{2.36}$$

where $\delta_{\boldsymbol{\Theta}_{1:t}^{(i)}}(.)$ is the Dirac function centered at $\boldsymbol{\Theta}_{1:t}^{(i)}$. Furthermore, for any measurable $\phi(\boldsymbol{\Theta}_{1:t})$ (as in Equation (2.31)) and $\forall t$ [36, 26, 20]:

$$\sum_{i=1}^{n} W^{(i)} \, g(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{1:t}) \rightarrow \int \phi(\boldsymbol{\Theta}_{1:t}) \; p(\boldsymbol{\Theta}_{1:t} \mid \boldsymbol{y}_{1:t}, \boldsymbol{\Theta}) d\boldsymbol{\Theta}_{1:t} \quad \textit{(almost surely)} \tag{2.37}$$

The particles are constructed recursively in time using a sequence of importance sampling densities $q_t(\boldsymbol{\Theta}_t \mid \boldsymbol{\Theta}_{t-1}, \boldsymbol{y}_t, \boldsymbol{\Theta})$. The importance weights are determined from the fact that:

$$p(\boldsymbol{\Theta}_{1:t} \mid \boldsymbol{y}_{1:t}, \boldsymbol{\Theta}) = p(\boldsymbol{\Theta}_{1:t-1} \mid \boldsymbol{y}_{1:t-1}, \boldsymbol{\Theta}) \frac{p(\boldsymbol{\Theta}_t \mid \boldsymbol{\Theta}_{t-1}, \boldsymbol{\Theta}) p(\boldsymbol{y}_t \mid \boldsymbol{\Theta}_t, \boldsymbol{\Theta})}{p(\boldsymbol{y}_t \mid \boldsymbol{y}_{1:t-1}, \boldsymbol{\Theta})} \tag{2.38}$$

Let $\{W^{(i)}, \boldsymbol{\Theta}_{1:t-1}^{(i)}\}_{i=1}^{N}$ the particulate approximation of $p(\boldsymbol{\Theta}_{1:t-1} \mid \boldsymbol{y}_{1:t-1}, \boldsymbol{\Theta})$. Note that for $t=1$ and for the Gaussian initial densities ν_0 of the proposed model, this consists of exact draws and weights $W^{(i)} = \frac{1}{N}$. At time t we proceed as follows ([45]):

1. Sample $\boldsymbol{\Theta}_t^{(i)} \sim q_t(\boldsymbol{\Theta}_t \mid \boldsymbol{\Theta}_{t-1}^{(i)}, \boldsymbol{y}_t, \boldsymbol{\Theta}), \forall i$ and set $\boldsymbol{\Theta}_{1:t}^{(i)} \leftarrow (\boldsymbol{\Theta}_{1:t-1}^{(i)}, \boldsymbol{\Theta}_t^{(i)})$
2. Compute incremental weights:

$$u_t^{(i)} = \frac{p(\boldsymbol{\Theta}_t^{(i)} \mid \boldsymbol{\Theta}_{t-1}^{(i)}, \boldsymbol{\Theta}) p(\boldsymbol{y}_t \mid \boldsymbol{\Theta}_t^{(i)}, \boldsymbol{\Theta})}{q_t(\boldsymbol{\Theta}_t^{(i)} \mid \boldsymbol{\Theta}_{t-1}^{(i)}, \boldsymbol{y}_t, \boldsymbol{\Theta})} \tag{2.39}$$

   and update the weights:

$$W^{(i)} = \frac{W^{(i)} u_t^{(i)}}{\sum_{j=1}^{N} W^{(i)} u_t^{(i)}} \tag{2.40}$$

3. Compute $ESS = \frac{1}{\sum_{i=1}^{N} (W^{(i)})^2}$ and if $ESS < ESS_{min}$ perform multinomial resampling to obtain a new population with equally weighted particles ($ESS_{min} = N/2$ was used in this study). Set $t \leftarrow t+1$ and go to step 1.

It can be easily established ([46]) that the locally optimal importance sampling density is:

$$q_t^{opt}(\boldsymbol{\Theta}_t \mid \boldsymbol{\Theta}_{t-1}, \boldsymbol{y}_t, \boldsymbol{\Theta}) = \frac{p(\boldsymbol{\Theta}_t \mid \boldsymbol{\Theta}_{t-1}, \boldsymbol{\Theta}) p(\boldsymbol{y}_t \mid \boldsymbol{\Theta}_t, \boldsymbol{\Theta})}{\int p(\boldsymbol{\Theta}_t \mid \boldsymbol{\Theta}_{t-1}, \boldsymbol{\Theta}) p(\boldsymbol{y}_t \mid \boldsymbol{\Theta}_t, \boldsymbol{\Theta}) \, d\boldsymbol{\Theta}_t} \tag{2.41}$$

In practice, it is usually impossible to sample from q_t^{opt} and/or calculate the integral in the denominator. As a result, approximations are used which nevertheless result in non-zero variance in the estimators. In this paper we take advantage of the fact that the transition density of $\boldsymbol{X}_t$ as well as the likelihood, conditioned on $\boldsymbol{z}_t$ are Gaussians and propose an importance sampling density of the form:

$$q_t(\boldsymbol{X}_t, \hat{\boldsymbol{z}}_t \mid \boldsymbol{X}_{t-1}, \hat{\boldsymbol{z}}_{t-1}, \boldsymbol{y}_t, \boldsymbol{\Theta}) = p(\hat{\boldsymbol{z}}_t \mid \hat{\boldsymbol{z}}_{t-1}, \boldsymbol{\Theta}) \frac{p(\boldsymbol{X}_t \mid \boldsymbol{X}_{t-1}, \boldsymbol{\Theta}) p(\boldsymbol{y}_t \mid \boldsymbol{X}_t, \boldsymbol{z}_t, \boldsymbol{\Theta})}{\int p(\boldsymbol{X}_t \mid \boldsymbol{X}_{t-1}, \boldsymbol{\Theta}) p(\boldsymbol{y}_t \mid \boldsymbol{X}_t, \boldsymbol{z}_t, \boldsymbol{\Theta}) \, d\boldsymbol{X}_t} \tag{2.42}$$

This implies using the prior to draw $\hat{\boldsymbol{z}}_t$ and the locally optimal distribution (conditioned on $\hat{\boldsymbol{z}}_t$ or equivalently $\boldsymbol{z}_t$) for $\boldsymbol{X}_t$. The latter will also be a Gaussian whose mean $\bar{\boldsymbol{\mu}}_t$ and covariance $\bar{\boldsymbol{S}}_X$ can be readily be established (e.g. using Kalman filter formulas):

$$\begin{aligned} \bar{\boldsymbol{S}}_X &= \left(\boldsymbol{S}_X^{-1} + \boldsymbol{W}_t^T \boldsymbol{\Sigma}^{-1} \boldsymbol{W}_t \right)^{-1} \\ \bar{\boldsymbol{\mu}}_t &= \bar{\boldsymbol{S}}_X \left(\boldsymbol{S}_X^{-1} \boldsymbol{\mu}_t + \boldsymbol{W}_t^T \boldsymbol{\Sigma}^{-1} \boldsymbol{y}_t \right) \end{aligned} \tag{2.43}$$

As a result the incremental weights u_t are given by:

$$u_t = \mid \bar{\boldsymbol{S}}_X \mid^{1/2} \exp\{ \frac{1}{2} \bar{\boldsymbol{\mu}}_t^T \bar{\boldsymbol{S}}_X^{-1} \bar{\boldsymbol{\mu}}_t - \frac{1}{2} \boldsymbol{\mu}_t^T \boldsymbol{S}_X^{-1} \boldsymbol{\mu}_t \} \tag{2.44}$$

2.4. Prediction and Bayesian adaptive time-integration. Bayesian inference results do not include just point estimates but rather samples from the posterior density, at least with repsect to the time-varying parameters $\boldsymbol{\Theta}_t$. The inferred posterior can be readily used to make *probabilistic predictions* about the future evolution of the high-dimensional, multiscale process $\boldsymbol{y}_t$. Given observations $\boldsymbol{y}_{1:\tau} = \{\boldsymbol{y}_t\}_{t=1}^{\tau}$, the *predictive posterior* for the future state of the system $\boldsymbol{y}_{\tau+1:\tau+T}$ over a time horizon T

can be expressed as:

$$
\begin{aligned}
p(\boldsymbol{y}_{\tau+1:\tau+T} \mid \boldsymbol{y}_{1:\tau}) &= \int p(\boldsymbol{y}_{\tau+1:\tau+T}, \boldsymbol{\Theta}, \boldsymbol{\Theta}_{\tau+1:\tau+T} \mid \boldsymbol{y}_{1:\tau})\ d\boldsymbol{\Theta} d\boldsymbol{\Theta}_{\tau+1:\tau+T} \\
&= \int \underbrace{p(\boldsymbol{y}_{\tau+1:\tau+T} \mid \boldsymbol{\Theta}, \boldsymbol{\Theta}_{\tau+1:\tau+T})}_{likelihood\ Equation(2.14)}\ p(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{\tau+1:\tau+T} \mid \boldsymbol{y}_{1:\tau}) d\boldsymbol{\Theta}\ d\boldsymbol{\Theta}_{\tau+1:\tau+T} \\
&= \int p(\boldsymbol{y}_{\tau+1:\tau+T} \mid \boldsymbol{\Theta}, \boldsymbol{\Theta}_{\tau+1:\tau+T}) \\
&\quad \underbrace{p(\boldsymbol{\Theta}_{\tau+1:\tau+T} \mid \boldsymbol{\Theta}_\tau, \boldsymbol{\Theta})}_{prior\ Equation(2.20)} \quad \underbrace{p(\boldsymbol{\Theta}, \boldsymbol{\Theta}_\tau \mid \boldsymbol{y}_{1:\tau})}_{posterior\ Equation(2.13)}\ d\boldsymbol{\Theta}\ d\boldsymbol{\Theta}_{\tau:\tau+T} \qquad (2.45)
\end{aligned}
$$

The integral above can be approximated using Monte Carlo. In particular given the particulate approximation of the posterior $p(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{1:\tau} \mid \boldsymbol{y}_{1:\tau})$ (which consists of samples of the dynamic variables $\boldsymbol{\Theta}_t$ and the MAP estimate of $\boldsymbol{\Theta}$), samples from the prior $p(\boldsymbol{\Theta}_{\tau+1:\tau+T} \mid \boldsymbol{\Theta}_\tau, \boldsymbol{\Theta})$ and subsequently the likelihood $p(\boldsymbol{\Theta}_{\tau+1:\tau+T} \mid \boldsymbol{\Theta}_\tau, \boldsymbol{\Theta})$ can readily be drawn. In fact, given that the latter is a multivariate Gaussian, the predictive posterior will consist of a mixture of Gaussians, one for each sample of $\boldsymbol{\Theta}_{\tau+1:\tau+T}$ drawn.

The important consequence of the Bayesian framework advocated is that precitive estimates are not restricted to point estimates but whole distributions which can readily quantify the predicitve uncertainty. This naturally gives rise to Bayesian, adaptive, time-integration scheme that allows bridging across timescales while providing quantitative, probabilistic estimates of the accuracy of the coarse-grained dynamics (Figure 2.5). The distribution of Equation (2.45) is used to probabilistically predict the evolution of the system. The time range over which the reduced model is employed does not have to be specified a priori but can be automatically determined by the variance of the predictive posterior (Figure 2.5). Once this exceeds the allowable tolerance specified by the analyst, the fine-scale process is reinitialized and more data are obtained, that can in turn be used to update the coarse-grained model. It is emphasized that due to the generative character of the model proposed, the reinitialization can be performed consistently based in general on the emission equations Equation (2.12). In contrast to existing techniques such as projective and coarse-projective integration [61, 63, 62, 60, 90, 114] as well as Heterogeneous Multiscale Methods [48, 101], there is no need to prescribe *lifting* and *restriction* operators and no ambiguity exists with regards to the appropriateness of the reinitialization scheme. Furthermore, the probabilistic coarse-grained model provides quantitative estimates for its predictive ability and automatically identifies the need for more information from the fine-scale model.

3. Numerical experiments. The first examples is intended to validate the accuracy of the proposed online EM scheme and utilizes a synthetic dataset. The second example uses actual data and illustrates the superiority of the proposed PMLDS model over existing SLDS models. Finally the third example provides an application in the context of multiscale simulations for the time-dependent diffusion equation.

3.1. Synthetic data. We generated data from the proposed model in order to investigate the ability of the inference and learning algorithms discussed. In particular, we considered a mixture of two $M = 2$, one-dimensional OU processes ($K = 1$) as in Equation (2.8) with $(b^{(1)}, \boldsymbol{q}_x^{(1)}, \boldsymbol{\Sigma}^{(1)}) = (0.1, -5.0, 0.2)$ (slow) and $(b^{(2)}, \boldsymbol{q}_x^{(2)}, \boldsymbol{\Sigma}^{(2)}) = (1., +5.0, 2.0)$ (fast). The logistic normal distribution was used to model the weights

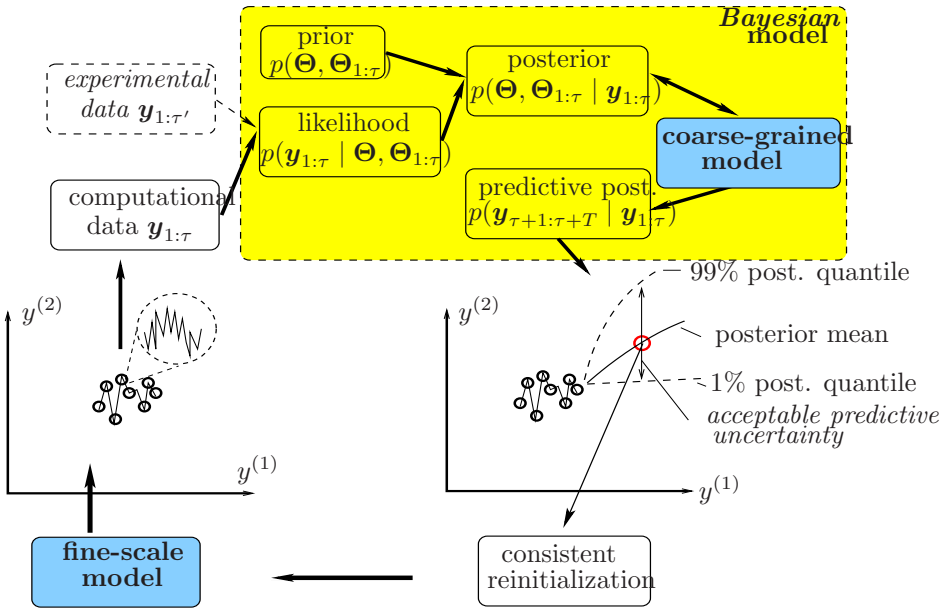
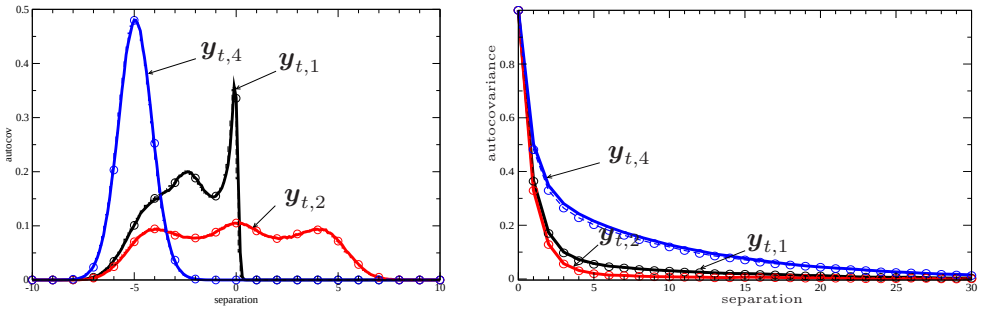


FIG. 2.5. Bayesian adaptive time-integration and data-model fusion illustrated for a two-dimensional flow. The data generated from computational simulations $\boldsymbol{y}_{1:\tau}$ and/or experiments $\boldsymbol{y}_{1:\tau'}$ are *sequentially* incorporated in the Bayesian model and the posterior $p(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{1:\tau} \mid \boldsymbol{y}_{1:\tau})$ over dynamic and static parameters is updated. The predictive posterior $p(\boldsymbol{y}_{\tau+1:\tau+T} \mid \boldsymbol{y}_{1:\tau})$ is over the time horizon T used to efficiently produce probabilistic predictions of the evolution of the high-dimensional process $\boldsymbol{y}_t$ in the future. When the uncertainty associated with those predictions exceeds the analysts' tolerance, the original system is *consistently* reinitialized and more data are generated. These are used to update the (predictive) posterior and to produce additional predictive estimates. It is noted that the tolerance in the predictive uncertainty can also be measured with respect to (low-dimensional) observables which are usually of interest in practical applications.

associated with each of the hidden processes using an isotropic Ornstein-Uhlenbeck process $d\mathbf{z}_t = -b_z(\mathbf{z}_t - \boldsymbol{q}_z)dt + \boldsymbol{\Sigma}_z^{1/2} d\boldsymbol{W}_t$ with $b_z = 1.0$, $\boldsymbol{q}_z = [0,\ 0]^T$ and $\boldsymbol{\Sigma}_z = \begin{bmatrix} 10. & 0 \\ 0 & 10. \end{bmatrix}$. Two 10×1 projection vectors $\boldsymbol{P}^m, m = 1, 2$ were generated from the prior $\mathcal{N}(\mathbf{0}, 100\boldsymbol{I})$ (see Appendix) and $(d = 10)$ time series $\boldsymbol{y}_t$ were produced based on Equation (2.12) with idiosyncratic variances $\boldsymbol{\Sigma} = 0.1^2\boldsymbol{I}$ and time step $\delta t = 1$. The resulting time series exhibit multimodal, non-Gaussian densities as can be seen in Figure 3.1(a) as well as two distinct time scales as it can be seen in the autocovariances plotted in Figure 3.1(b).

Figure 3.2 depicts the convergence of the proposed online EM scheme to the reference values of $b_x^{(m)}, m = 1, 2$, for various block sizes L and particle populations N. Figure 3.3 depicts the evolution of the log-likelihood per iteration of the EM algorithm. Figure 3.4(a) depicts the normalized error in the identified $\boldsymbol{P}^{(m)}, m = 1, 2$ and isiosyncratic variances $\boldsymbol{\Sigma}$ pre coordinate after 20,000 iterations. In all cases the algorithm exhibits good convergence to the reference values.

3.2. Temperature Dataset. The goal of this numerical experiment is to illustrate the interpretability of the proposed model and compare with the switching-state linear model discussed in section 2.1 (Equation (2.6)). For that purpose we utilized the temperature data (in degrees Fahrenheit) of the capitals of the 50 states in the U.S.A



(a) density (b) autocovariance

FIG. 3.1. Densities (a)) and Autocovariances (b)) of times series $\boldsymbol{y}_{t,1}$ (black), $\boldsymbol{y}_{t,2}$ (red) and $\boldsymbol{y}_{t,4}$ (blue) (solid lines). With $(-\circ-)$ the densities and autocovariances of the same times series generated using the learned model parameters using the proposed online EM scheme with $L = 200$ and $N = 200$ (see Figures 3.2 and 3.4)

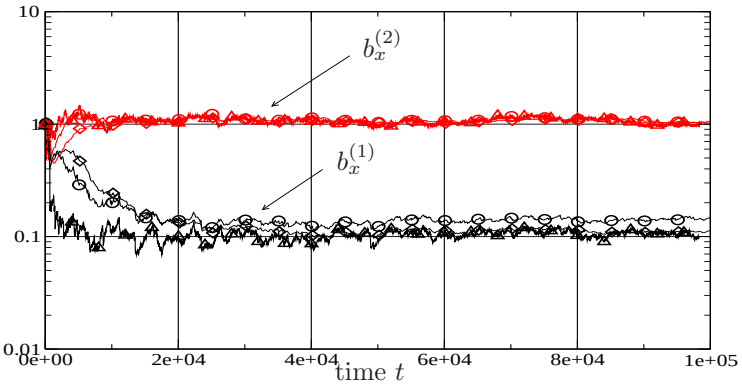


FIG. 3.2. *Convergence of* $b_x^{(1)}$ *(black) and* $b_x^{(2)}$ *(red) using the online EM algorithm for three different combinations of* L *and* N. $(-\circ-)$ *corresponds to* $L = 100$, $N = 100$, $(-\diamond-)$ *to* $L = 200$, $N = 200$ *and* $(-\triangle-)$ *to* $L = 20$, $N = 1000$

$(d = 50)$. The data was obtained from *http://www.engr.udayton.edu/weather/citylistUS.htm* and it represents the average daily temperatures from 01/01/1996 until 01/13/2009 (i.e. 5,127 daily observations).

Figure 3.5 depicts the posterior memberships corresponding to the SLDS and PMLDS models based on a reduced model with two hidden states ($M = 2$) described by one-dimensional OU processes ($K = 1$). The former assumes that at each time instant the observables $\boldsymbol{y}_t$ arises from a *single* hidden process. Hence a single entry of $\boldsymbol{z}_t = [z_{1,t}, z_{2,t}, \ldots, zM, t]$ is equal to 1 and the rest are all equal to 0. The top part of Figure 3.5 shows the posterior mean of $z_{m,t}, m = 1, 2$. As one would expect the two-states correspond to cold-winter (blue) and hot-summer (red) and alternate periodically (roughly the cold-winter state is active between early November until mid-April and the hot-summer state in the remainder of the calendar year). The top part of Figure 3.7 depicts the corresponding $\boldsymbol{P}^{(m)}, m = 1, 2$ where southern states have higher values and northern states lower. Naturally, winter and summer represent the

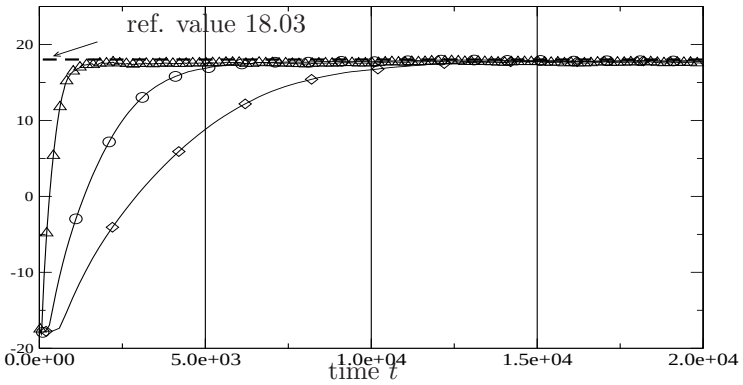
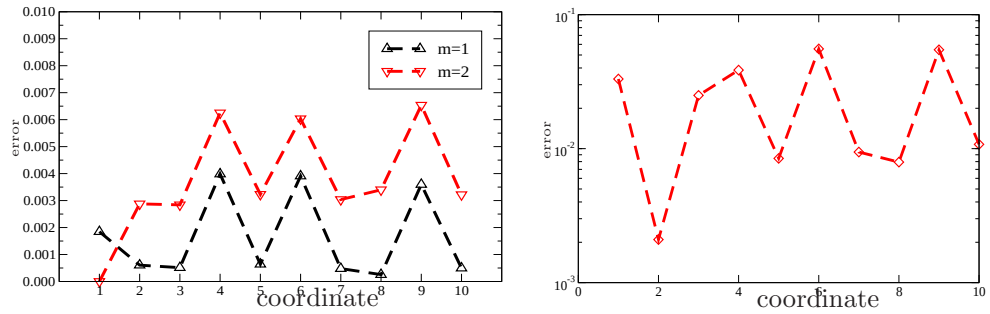


FIG. 3.3. *Log-likelihood convergence using the online EM algorithm for three different combinations of L and N. $(-\circ-)$ corresponds to $L = 100$, $N = 100$, $(-\Diamond-)$ to $L = 200$, $N = 200$ and $(-\triangle-)$ to $L = 20$, $N = 1000$*



(a) normalized absolute error in $\boldsymbol{P}_j^{(m)}$

(b) normalized abs. error in σ_j^2

FIG. 3.4. Normalized absolute errors (per coordinate) on the identification of the projection vectors $\boldsymbol{P}^{(m)}, m = 1, 2$ and idiosyncratic variances $\sigma_j^2, j = 1, \ldots, d$ using the proposed online EM scheme with $L = 200$ and $N = 200$

two extremes but several intermediate states are also present. The proposed partial-membership model can account for those states without increasing the cardinality of the reduced-order dynamics. As it can be seen in the bottom part of Figure 3.5 which depicts the particulate approximation of the posterior of $z_{m,t}, m = 1, 2$, the two hidden states can also be attributed to the two extremes but the actual temperatures arise by a weighted combination of these two. Naturally during spring-summer months the weight of the "red" state is higher and during autumn-winter months the weight of the "blue" state takes over. The posterior of the hidden processes $\boldsymbol{x}_t^{(m)}, m = 1, 2$ is depicted in Figure 3.6.

In order to quantitatively compare the two models we calculated the average,

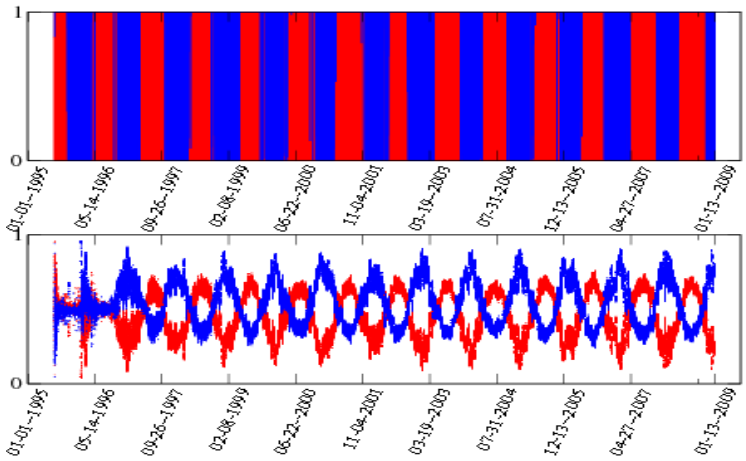


FIG. 3.5. *(Top) Posterior mean of $z_{m,t}, m = 1, 2$ based on the SLDS and (Bottom) particulate approximation of the posterior of $z_{m,t}, m = 1, 2$ PMLDS. Both results were obtained using the previously discussed online EM scheme with $L = 200$ and $N = 100$*

M	K	SLDS	PMLDS
2	1	-179.97 ± 37.31	-171.11 ± 37.20
2	2	-170.68 ± 36.95	-141.11 ± 27.82
4	1	-176.40 ± 34.36	-143.81 ± 25.56
4	2	-166.05 ± 30.57	-117.67 ± 21.15

TABLE 3.1

One-step-ahead predictive log-likelihood (Equation (3.1)) of SLDS and PMLDS models for various M, K. The table reports the average value plus/minus one standard deviation in bits. All results were obtained using the previously discussed online EM scheme with $L = 200$ and $N = 100$

one-step-ahead, predictive log-likelihood $\log p(\boldsymbol{y}_{t+1} \mid \boldsymbol{y}1 : t), \forall t \in [0, T)$:

$$\begin{aligned} p(\boldsymbol{y}_{t+1} \mid \boldsymbol{y}_{1:t}) &= \int \log p(\boldsymbol{y}_{t+1} \mid \boldsymbol{\Theta}, \boldsymbol{\Theta}_{t+1} \boldsymbol{y}_{1:t}) \; p(\boldsymbol{\Theta}, \boldsymbol{\Theta}_{t+1} \mid \boldsymbol{y}_{1:t}) d\boldsymbol{\Theta} d\boldsymbol{\Theta}_{t+1} \\ &= \int \log p(\boldsymbol{y}_{t+1} \mid \boldsymbol{\Theta}, \boldsymbol{\Theta}_{t+1} \boldsymbol{y}_{1:t}) \\ &\quad \underbrace{p(\boldsymbol{\Theta}_{t+1} \mid \boldsymbol{\Theta}_t, \boldsymbol{\Theta})}_{prior} \; \underbrace{p(\boldsymbol{\Theta}, \boldsymbol{\Theta}_t \mid \boldsymbol{y}_{1:t})}_{posterior} d\boldsymbol{\Theta} \; d\boldsymbol{\Theta}_{t:t+1} \end{aligned} \tag{3.1}$$

The latter integral is approximated by Monte Carlo using the MAP estimate of $\boldsymbol{\Theta}$ and the particulate approximation of the posterior for the dynamic variables $\boldsymbol{\Theta}_t$. This provides a measure of how well the model generalizes to a novel observation from the same distribution as the training data and higher values imply a better model. Table 3.1 reports the average values (in bits) plus/minus the standard deviation (over $t \in (100, T = 5127)$). Similar calculations were carried out for other model cardinalities (i.e. M, K) and in all cases the proposed model exhibited superior performance. This superiority becomes more pronounced as M and K increased which can be attributed to the factorial character of PMLDS.

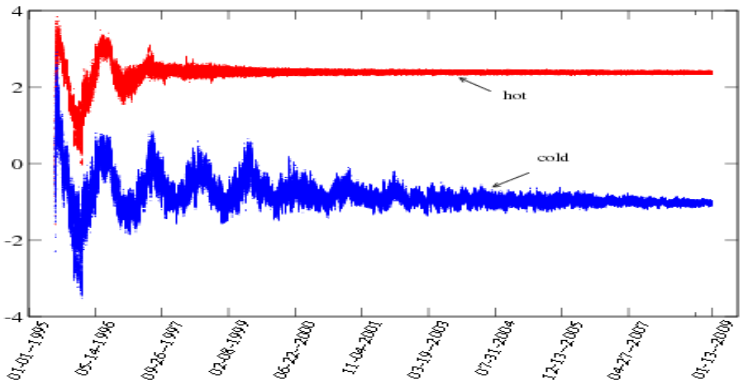


FIG. 3.6. *Particulate approximation of the posterior of* $\boldsymbol{x}_t^{(m)}, m = 1, 2$. *Results were obtained*

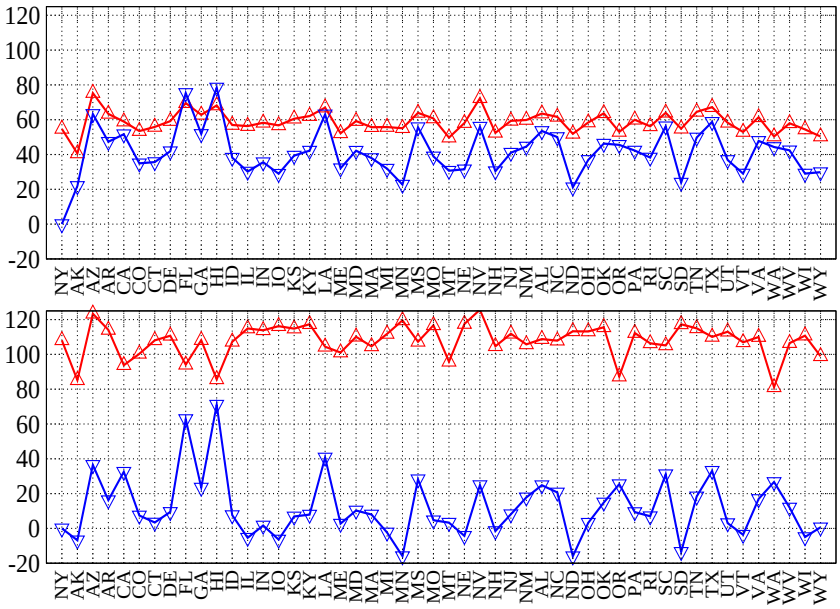


FIG. 3.7. *Maximum posterior estimates of* $\boldsymbol{P}^{(m)} \in \mathbb{R}^{50}, m = 1, 2$ *based on the SLDS(top) and PMLDS (bottom) models. Both results were obtained using the previously discussed online EM scheme with* $L = 200$ *and* $N = 100$

3.3. Transient Heat equation. We finally apply the proposed analysis scheme to the one-dimensional transient heat equation:

$$\begin{cases} \frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(a(x)\frac{\partial u}{\partial x}\right), & x \in [0,1] \\ u(0,t) = u(1,t) = 0, & \forall t \end{cases} \tag{3.2}$$

The spatial domain was discretized with $1,000$ finite elements of equal length and we considered a "rough" conductivity profile shown in Figure 3.8(a). The conductivity $a(x)$ in each finite element was assumed constant and its value was drawn indepen-

dently from a uniform distribution [2]. For $x \in [0, 0.5]$ we used the uniform $U[0.01, 0.1]$ and for $x \in (0.5, 1]$, $U[0.51, 0.6]$. This naturally resulted in the jump observed in Figure 3.8(a) which as a consequence gave rise to two distinct slow time scales in the solution profile $u(x,t)$ depicted in Figure 3.8(b). A rough profile of initial conditions was also used (as can be seen Figure 3.8(b), $t = 0$). In particular at each node $x_i = 0.001i,\ i = 1, \ldots, 1001$ we set $u(x_i, 0) = 10x_i(1 - x_i)(1 + 0.1Z_i)$ where $Z_i \sim N(0, 1)$ (i.i.d).

Upon spatial discretization, we obtain a coupled system of ODEs:

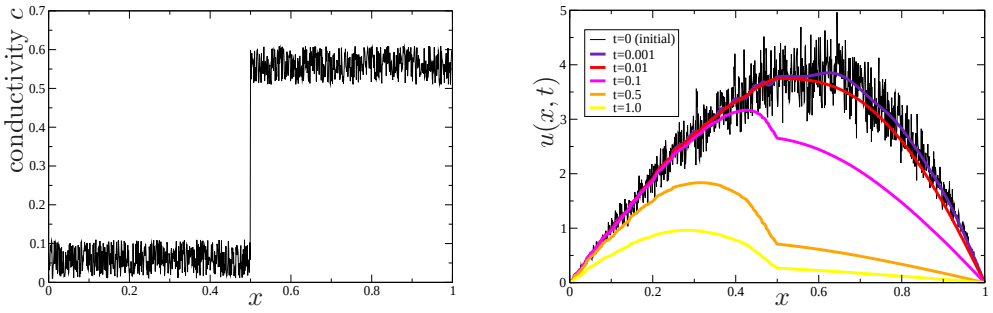
$$\dot{\boldsymbol{y}}_t + \boldsymbol{K}\boldsymbol{y}_t = \boldsymbol{0} \tag{3.3}$$

where $\boldsymbol{y}_t \in \mathbb{R}^{1001}$ represents the solution at the nodes x_i, i.e. $\boldsymbol{y}_t = [u(x_1, t), u(x_2, t), \ldots, \ldots, u(x_d, t)]^T$. In contrast to existing approaches for the same diffusion equation (e.g. [2, 1, 118]) we do not exploit mathematical properties of the PDE in specifying the coarse-grained model but rely on data. This data is obtained upon temporal discretization of Equation (3.3) where a time step $\delta t = 0.0001$ was used. As a result at each time step we obtained a vector of observables $\boldsymbol{y}_t$ of dimension $d = 1001$. This data was incorporated in the Bayesian model proposed using two hidden OU process ($M = 2$) of dimension $K = 2$ each. In particular we employed the online EM scheme previously discussed over blocks of length $L = 10$ time steps and $N = 100$ particles. In particular (see also Figure 2.5):

- data over 20 times steps δt, $\boldsymbol{y}_{1:20}$ (i.e. corresponding to total time $20\delta t = 0.002$) were ingested by the Bayesian reduced model, and
- the latter was used to predict the evolution of the system over 500 time steps (i.e. total time $T = 500\delta t = 0.05$).
- The original solver of the governing PDE was then re-initialized using the posterior mean estimate of the state of the system $\boldsymbol{y}_{520}$ and was run for further 20 time steps. Using the additional data obtained $\boldsymbol{y}_{521:540}$, the Bayesian model was updated and the process described was repeated.

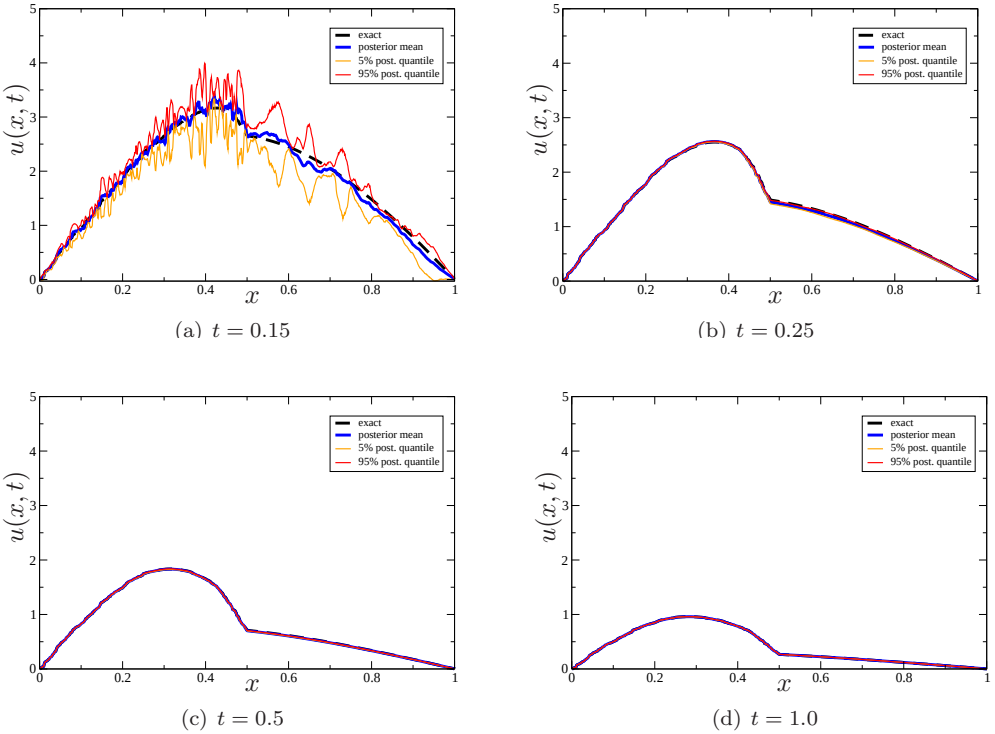
It is noted that the proposed Bayesian prediction scheme results in a reduction of the number of fine scale integration time steps by a factor of 25 ($T/20\delta t = 0.05/0.002$) leading to a significant acceleration of the simulation process. Figure 3.9 depicts the posterior estimates of the solution at various time instants. In all cases these approximate very accurately the exact solution and these estimates improve as more are accumulated. One of the main advantages of the proposed approach is that apart from single-point estimates one can readily obtain credible intervals that quantify predictive uncertainties due to information loss by the use of the reduced-order dynamic model and the finite amount of data used to learn that model. As it is seen in Figure 3.9 these envelop the exact solution and become tighter at larger times. As one would expect, when a larger predictive horizon $T = 0.1$ (i.e. $1,000$ time steps δt) is used, as it can be seen in Figures 3.10 and 3.11 the predictive uncertainty grows. Such a scheme however is twice as efficient leading to a reduction of computational effort by a factor of 50 (i.e. $T/20\delta t = 0.1/0.002$). Hence if the analyst is willing to tolerate the additional uncertainty, efficiency gains can be achieved. This supports the arguments made previously with regards to an *adaptive* Bayesian scheme where, the level of predictive uncertainty would be specified and the algorithm would automatically revert to the fine scale model in order to obtain more data and improve the predictive estimates.

[2] We considered a single realization of the conductivity profile and solved for it as a deterministic problem. The stochastic PDE where $a(x)$ is random is not considered here.



(a) conductivity (b) Exact solution profiles $u(x,t)$ at various t

FIG. 3.8. Dynamic heat equation. Spatial discretization with 1,000 finite elements. Time discretization using $\delta t = 1 \times 10^{-4}$.

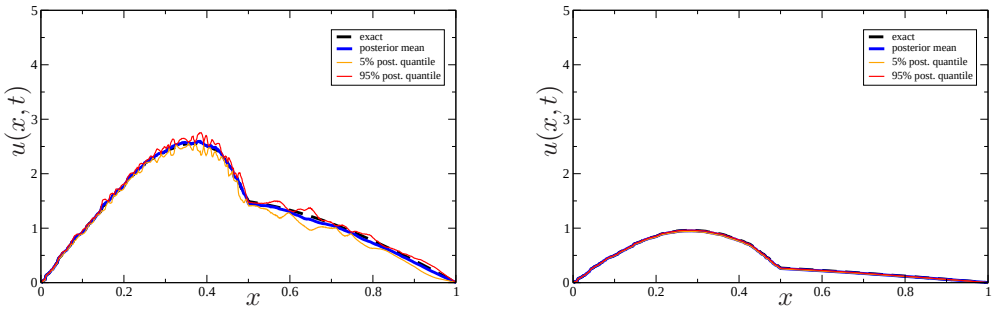


(a) $t = 0.15$ (b) $t = 0.25$

(c) $t = 0.5$ (d) $t = 1.0$

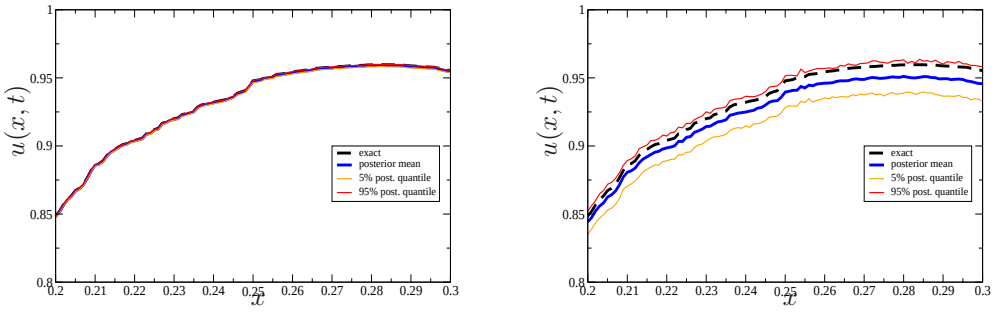
FIG. 3.9. Comparison of predictive posterior estimates (posterior mean and 5% and 95% quantiles) with exact solution $u(x,t)$ at various t

4. Conclusions. The proposed modeling framework can extract interpretable reduced representations of high-dimensional systems by employing simple, low-dimensional processes. It simultaneously achieves dimensionality reduction and learning of reduced dynamics. The Bayesian framework adopted provides a generalization over single-point estimates obtained through maximum-likelihood procedures. It can quantify uncertainties associated with learning from finite amounts of data and produce probabilistic predictive estimates. The latter can be used to rigorously perform concurrent



(a) $t = 0.25$ (b) $t = 1.0$

FIG. 3.10. Comparison of predictive posterior estimates (posterior mean and 5% and 95% quantiles) with exact solution $u(x,t)$ at various t. These results were obtained with a prediction horizon $T = 0.1$ ($\delta t = 0.0001$) in contrast to Figure 3.9 which were obtained for $T = 0.05$.



(a) $T = 0.05$ (b) $T = 0.1$

FIG. 3.11. Comparison of predictive posterior estimates (posterior mean and 5% and 95% quantiles) with exact solution $u(x,t)$ at $t = 1.0$. These results were obtained with a prediction horizon (a) $T = 0.05$ and (b) $T = 0.1$.

simulations with the microscopic model without the need of prescribing ad hoc upscaling and downscaling operators.

Critical to the efficacy of the proposed techniques is scalability particularly with regards to the large dimension d of the original process. The algorithms proposed imply $O(d)$ order of operations. Furthermore they can dynamically *update* the coarse-grained models as more data become available. In a typical scenario, the fine-scale model is reinitialized several times in order to obtain additional information about the system's evolution that is incorporated in the coarse-grained dynamics "on the fly".

The Bayesian, statistical perspective can readily be extended to the modeling *stochastic* dynamical systems. This would require generating more than one realizations of the original dynamics which can nevertheless be incorporated in the coarse-grained models using the same online EM scheme. In fact the loss of information that unavoidably takes place during the coarse-graining, results in probabilistic reduced-order models even if the original model was deterministic. A critical question that offers opportunity for future research on the topic relates to structural learning and in particularly with the dimensionality of the representation, i.e. the number of hid-

den processes $\boldsymbol{x}_t^{(m)}$ needed (denoted by M in Equation (2.7)). Treating this as a model selection problem as it was done in the examples, assumes that there is a single, optimal finite-dimensional representation. Current research activities are centered around *nonparametric* Bayesian priors over infinite combinatorial structures based on the Dirichlet process paradigm and infinite latent features models (e.g. [129, 70, 75]). These offer an alternative perspective by assuming that the number of building blocks is potentially unbounded, and that the observables only manifest a sparse subset of those. As a result, the *cardinality* of the coarse-grained model can be automatically determined from the data. Another aspect that warrants further investigation is prior modeling of the static parameters. Apart from the regularization effect this offers, it can promote the discovery of desirable features, such as slow-varying essential dynamics, sparse factors (e.g. $\boldsymbol{P}^{(m)}$ in Equation (2.12)) which can advance the interpretability of the results and facilitate the inference process.

Appendix. This appendix discusses the sufficient statistics and update equations for the static parameters $\boldsymbol{\Theta}$ used in the probabilistic model proposed. In the first section we discuss parameters appearing in the reduced-order dynamics models and in the second those appearing in the likelihood.

Sufficient statistics for parameters appearing in the prior. As discussed in section 2.2.1, independent, isotropic OU processes are used as prior models for the latent, coarse-grained dynamics $\boldsymbol{x}_t^{(m)} \in \mathbb{R}^K$ as well the process $\boldsymbol{z}_t \in \mathbb{R}^M$ that models the frctional memberships to to each process m. We therefore discuss the essential elements for the online EM computations described in section 2.3 ([6, 7, 8]) for a general isotropic OU process in $\mathbb{R}^n$ of the form:

$$d\boldsymbol{x}_t = -b(\boldsymbol{x}_t - \boldsymbol{q})dt + \boldsymbol{S}^{1/2} d\boldsymbol{W}_t \tag{4.1}$$

It is of interest to determine the parameters $\boldsymbol{\theta} = (b, \boldsymbol{q}, \boldsymbol{S})$. Let also $\pi(\boldsymbol{\theta})$ denote the prior on $\boldsymbol{\theta}$. The readers can adjust the expressions below to any $\boldsymbol{x}_t^{(m)}$ or $\boldsymbol{z}_t$ since independent priors were used. Note that the stationary distribution of $\boldsymbol{x}_t$ is a Gaussian:

$$\nu_0(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{q}, \boldsymbol{C} = \frac{1}{2b}\boldsymbol{S}) \tag{4.2}$$

and the transition density $p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$ assuming that equidistant time instants with time step δt are considered, is given by:

$$p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t \mid \boldsymbol{\mu}_{\delta t}(\boldsymbol{x}_{t-1}), \boldsymbol{S}_{\delta t}) \tag{4.3}$$

where:

$$\boldsymbol{\mu}_{\delta t}(\boldsymbol{x}_{t-1}) = \boldsymbol{x}_{t-1} - (1 - e^{-b\delta t})(\boldsymbol{x}_{t-1} - \boldsymbol{q}) \tag{4.4}$$

and:

$$\boldsymbol{S}_{\delta t} = \frac{1 - e^{-2b\delta t}}{2b}\boldsymbol{S} \tag{4.5}$$

Given a block of length L with observables $\boldsymbol{y}_{1:L}$ and according to Equations (2.27) and (2.28) we have that:

$$\begin{aligned} \hat{Q}(\boldsymbol{\theta}^{(k-1)}, \boldsymbol{\theta}) \quad &= \big\langle -\tfrac{1}{2}\log \mid \boldsymbol{C} \mid -\tfrac{1}{2}(\boldsymbol{x}_1 - \boldsymbol{q})^T \boldsymbol{C}^{-1}(\boldsymbol{x}_1 - \boldsymbol{q}) \\ &\quad + \textstyle\sum_{t=2}^{L} -\tfrac{1}{2}\log \mid \boldsymbol{S}_{\delta t} \mid -\tfrac{1}{2}(\boldsymbol{x}_t - \boldsymbol{q})^T \boldsymbol{C}^{-1}(\boldsymbol{x}_1 - \boldsymbol{q}) \big\rangle \\ &\quad + \log \pi(\boldsymbol{\theta}) \end{aligned}$$

where the brackets $< . >$ imply expectation with respect to $\hat{p}(\boldsymbol{x}_{1:L} \mid \boldsymbol{\theta}^{(k-1)}, \boldsymbol{y}_{1:L})$ as in Equation (2.28). In order to maximize $\hat{Q}(\boldsymbol{\Theta}^{(1:k-1)}, \boldsymbol{\Theta})$ as in Equation (2.30) one needs to solve the system of equations arising from $\frac{\partial \hat{Q}(\boldsymbol{\theta}^{(1:k-1)}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \boldsymbol{0}$ These equations equations with respect to $\boldsymbol{\theta}$ are solved with fixed point iterations. They depend on

the following 7 sufficient statistics $\boldsymbol{\Phi} = \{\Phi_j\}_{j=1}^7$:

$$\begin{aligned}
\Phi_1 &= < \boldsymbol{x}_1 > \\
\Phi_2 &= < \boldsymbol{x}_1 \boldsymbol{x}_1^T > \\
\Phi_3 &= \left\langle \sum_{t=2}^L \boldsymbol{x}_{t-1} \right\rangle \\
\Phi_4 &= \left\langle \sum_{t=2}^L \boldsymbol{x}_t - \boldsymbol{x}_{t-1} \right\rangle \\
\Phi_5 &= \left\langle \sum_{t=2}^L \boldsymbol{x}_{t-1} \boldsymbol{x}_{t-1}^T \right\rangle \\
\Phi_6 &= \left\langle \sum_{t=2}^L (\boldsymbol{x}_t - \boldsymbol{x}_{t-1}) \boldsymbol{x}_{t-1}^T \right\rangle \\
\Phi_7 &= \left\langle \sum_{t=2}^L (\boldsymbol{x}_t - \boldsymbol{x}_{t-1})(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})^T \right\rangle
\end{aligned} \tag{4.6}$$

Sufficient statistics for parameters appearing in the likelihood. The process a bit more involved in the case of the parameters appearing in the likelihood Equation (2.14) i.e. the projection matrices $\{\boldsymbol{P}^{(m)}\}_{m=1}^M$ of dimension $d \times K$ and the covariance $\boldsymbol{\Sigma}$ which is a (positive definite) matrix of $d \times d$. In order to retain scalability in high-dimensional problems (i.e. $d >> 1$) we assume a diagonal form of $\boldsymbol{\Sigma} = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2)$ which implies learning d parameters rather than $d(d+1)/2$.

Denoting now by $\boldsymbol{\theta} = (\{\boldsymbol{P}^{(m)}\}_{m=1}^M, \{\sigma_j^2\}_{j=1}^d)$, $\pi(\boldsymbol{\theta})$ the prior and according to Equations (2.27) and (2.28) we have that:

$$\begin{aligned}
\hat{Q}(\boldsymbol{\theta}^{(k-1)}, \boldsymbol{\theta}) \quad &= \left\langle \sum_{t=1}^L -\tfrac{1}{2} \log \mid \boldsymbol{\Sigma} \mid -\tfrac{1}{2} (\boldsymbol{y}_t - \boldsymbol{W}_t \boldsymbol{X}_t)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{y}_t - \boldsymbol{W}_t \boldsymbol{X}_t) \right\rangle \\
&+ \log \pi(\boldsymbol{\theta})
\end{aligned} \tag{4.7}$$

Differentiation with respect to $\boldsymbol{P}^{(m)}$ reveals that the stationary point must satisfy:

$$\boldsymbol{A}^{(m)} = \sum_{n=1}^M \boldsymbol{P}^{(n)} \boldsymbol{B}^{(n,m)} \tag{4.8}$$

where the sufficient statistics are:

$$\underbrace{\boldsymbol{A}^{(m)}}_{d \times K} = \left\langle \sum_{t=1}^L z_{t,m} \boldsymbol{y}_t (\boldsymbol{x}_t^{(m)})^T \right\rangle, \quad m = 1, 2, \ldots, M \tag{4.9}$$

and:

$$\underbrace{\boldsymbol{B}^{(n,m)}}_{K \times K} = \left\langle \sum_{t=1}^L z_{t,n} z_{t,m} \boldsymbol{x}_t^{(n)} (\boldsymbol{x}_t^{(m)})^T \right\rangle \tag{4.10}$$

In the absence of a prior $\pi(\boldsymbol{\theta})$ and if $\boldsymbol{P}_j^{(m)}$ and $\boldsymbol{A}_j^{(m)}$ represent the j^{th} rows ($j = 1, \ldots, d$) of the matrices $\boldsymbol{P}^{(m)}$ and $\boldsymbol{A}^{(m)}$ respectively, then Equation (4.9) implies:

$$\begin{aligned}
\underbrace{\begin{bmatrix} \boldsymbol{A}_j^{(1)} & \boldsymbol{A}_j^{(2)} & \ldots & \boldsymbol{A}_j^{(M)} \end{bmatrix}}_{\boldsymbol{A}_j : (1 \times K \;\; M)} = \quad & \underbrace{\begin{bmatrix} \boldsymbol{P}_j^{(1)} & \boldsymbol{P}_j^{(2)} & \ldots & \boldsymbol{P}_j^{(M)} \end{bmatrix}}_{\boldsymbol{P}_j : (1 \times K \;\; M)} \\
& \underbrace{\begin{bmatrix} \boldsymbol{B}^{(1,1)} & \boldsymbol{B}^{(1,2)} & \ldots & \boldsymbol{B}^{(1,M)} \\ \boldsymbol{B}^{(2,1)} & \boldsymbol{B}^{(2,2)} & \ldots & \boldsymbol{B}^{(2,M)} \\ \ldots & \ldots & \ldots & \ldots \\ \boldsymbol{B}^{(M,1)} & \boldsymbol{B}^{(M,2)} & \ldots & \boldsymbol{B}^{(M,M)} \end{bmatrix}}_{\boldsymbol{B} : (MK \times MK)}
\end{aligned} \tag{4.11}$$

This leads to the following update equations for $\boldsymbol{P}_j^{(m)},\ \forall j,m$:

$$\boldsymbol{P}_j = \boldsymbol{A}_j \boldsymbol{B}^{-1} \tag{4.12}$$

Note that the matrix $\boldsymbol{B}$ to be inverted is independent of the dimension of the observables d ($d >> 1$) and the inversion needs to be carried out once for all $j = 1,\ldots,d$. Hence the *scaling of the update equations for* $\boldsymbol{P}^{(m)}$ *is* $O(d)$ i.e. linear with respect to the dimensionality of the original system.

Furthermore, in the absence of a prior $\pi(\boldsymbol{\theta})$, differentiation with respect to σ_j^{-2} ($j = 1,\ldots,d$) leads to the following update equation:

$$L\ \sigma_j^2 \quad = \sum_{t=1}^{L} y_{t,j}^2 - 2\boldsymbol{A}_j\ \boldsymbol{P}_j^T + \boldsymbol{P}_j\ \boldsymbol{B}\boldsymbol{P}_j^T \tag{4.13}$$

In summary the sufficient statistics needed are the ones in Equations (4.9) and (4.12).

In the numerical examples in this paper a diffuse Gaussian prior was used for $\boldsymbol{P}^{(m)}$ with variance 100 for each of the entries of the matrix. This leads to the addition of the term $1/100$ in the diagonal elements of the $\boldsymbol{B}$ in Equation (4.12). No priors were used for σ_j^2.

REFERENCES

[1] A. Abdulle and B. Engquist, *Finite element heterogeneous multiscale methods with near optimal computational complexity*, MULTISCALE MODELING & SIMULATION, 6 (2007), pp. 1059 – 1084.

[2] A. Abdulle and E. Weinan, *Finite difference heterogeneous multi-scale method for homogenization problems*, JOURNAL OF COMPUTATIONAL PHYSICS, 191 (2003), pp. 18 – 39.

[3] F.F. Abraham, R. Walkup, H.J. Gao, M. Duchaineau, T.D. De la Rubia, and M. Seager, *Simulating materials failure by using up to one billion atoms and the world's fastest computer: Work-hardening*, PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, 99 (2002), pp. 5783 – 5787.

[4] Charu C. Aggarwal, *A framework for clustering massive-domain data streams*, in ICDE, 2009, pp. 102–113.

[5] O. Aguilar and M. West, *Bayesian dynamic factor models and variance matrix discounting for portfolio allocation*, Journal of Business and Economic Statistics, 18 (2000), p. 338357.

[6] C. Andrieu and A. Doucet, *Online expectation-maximization type algorithms for parameter estimation in general state space models*, in IEEE International Conference on Acoustics, Speech, and Signal Processing, 6-10 April 2003, pp. 69–72.

[7] C. Andrieu, A. Doucet, S.S. Singh, and V.B. Tadić, *Particle methods for change detection, identification and control*, in Proceedings of the IEEE, vol. 92, 2004, pp. 423–438.

[8] C. Andrieu, A. Doucet, and V.B. Tadić, *Online simulation-based methods for parameter estimation in non linear non gaussian state-space models*, in Proc. IEEE CDC (invited paper), 2005.

[9] Arik Azran and Zoubin Ghahramani, *Spectral methods for automatic multiscale data clustering*, in CVPR (1), 2006, pp. 190–197.

[10] F.R. Bach and M.I. Jordan, *Learning spectral clustering, with application to speech separation*, JOURNAL OF MACHINE LEARNING RESEARCH, 7 (2006), pp. 1963 – 2001.

[11] Gökhan H. Bakir, Alexander Zien, and Koji Tsuda, *Learning to find graph pre-images*, in DAGM-Symposium, 2004, pp. 253–261.

[12] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen, *The infinite hidden markov model*, in Neural Information Processing Systems 14, T.G. Dietterich, S. Becker, and Z. Ghahramani, eds., MIT Press, 2002, pp. 577–585.

[13] C. Bishop, *Latent variable models*, in Learning in Graphical Models, M. I. Jordan, ed., MIT Press, 1999, pp. 371–403.

[14] D Blei, T Griffiths, M Jordan, and J Tenenbaum, *Hierarchical topic models and the nested chinese restaurant process*, in NIPS 2003, 2003.

[15] D. Blei and J. Lafferty, *Dynamic topic models*, in 23rd International Conference on Machine Learning, p. 2006.

[16] D. Blei, A. Ng, and M. Jordan, *Latent dirichlet allocation*, Journal of Machine Learning Research, (2003), pp. 993–1022.

[17] E. Cances, F. Legoll, and G. Stoltz, *Theoretical and numerical comparison of some sampling methods for molecular dynamics*, Mathematical Modelling and Numerical Analysis, 41 (2007), p. 351.

[18] O. Cappé, *Ten years of HMMs (An HMM bibliography)*. 2001.

[19] Olivier Cappé and Eric Moulines, *Online em algorithm for latent data models*, CoRR, abs/0712.4273 (2007).

[20] O. Cappé, E. Moulines, and T. Rydén, *Inference in Hidden Markov Models*, Springer-Verlag, 2005.

[21] F. Caron, M. Davy, and A. Doucet, *Generalized Polya urn for time-varying Dirichlet processes* , in Proceedings of Uncertainty in Artificial Intelligence 2007, 2007.

[22] L. Chacón, *Scalable parallel implicit solvers for 3d magnetohydrodynamics*, Journal of Physics: Conference Series, 125 (2008), p. 012041.

[23] Wassim S. Chaer, Robert H. Bishop, and Joydeep Ghosh, *A mixture-of-experts framework for adaptive kalman filtering*, IEEE Transactions on Systems, Man, and Cybernetics, Part B, 27 (1997), pp. 452–464.

[24] C. B. Chang and M. Athans, *State estimation for discrete systems with switching parameters*, IEEE Transactions on Aerospace and Electronic Systems, AES, 14, p. 418424.

[25] C. Chipot and A. Pohorille, eds., *Free energy calculations*, Springer Series in Chemical Physics, 2007.

[26] N Chopin, *Central limit theorem for sequential monte carlo methods and its application to bayesian inference*, ANNALS OF STATISTICS, 32 (2004), pp. 2385 – 2411.

[27] A.J. Chorin, O.H. Hald, and R. Kupferman, *Optimal prediction and the Mori-Zwanzig representation of irreversible processes*, PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, 97 (2000), pp. 2968 – 2973.

[28] ———, *Prediction from partial data, renormalization, and averaging*, JOURNAL OF SCIENTIFIC COMPUTING, 28 (2006), pp. 245 – 261.

[29] A.J. Chorin and P. Stinis, *Problem reduction, renormalization, and memory*, Comm. Appl. Math. Comp. Sc., 1 (2005), pp. 1–27.

[30] G. Ciccotti, R. Kapral, and A. Sergi, *Non-equilibrium molecular dynamics*, in Handbook of materials modeling, S. Yip, ed., 2005, pp. 745–761.

[31] R.R. Coifman, I.G. Kevrekidis, S. Lafon, M. Maggioni, and B. Nadler, *Diffusion maps, reduction coordinates, and low dimensional representation of stochastic systems*, MULTISCALE MODELING & SIMULATION, 7 (2008), pp. 842 – 864.

[32] R.R. Coifman, S. Lafon, A.B. Lee, M. Maggioni, B. Nadler, F. Warner, and S.W. Zucker, *Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps*, PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, 102 (2005), pp. 7426 – 7431.

[33] Charles D. Cranor, Theodore Johnson, Oliver Spatscheck, and Vladislav Shkapenyuk, *Gigascope: A stream database for network applications*, in SIGMOD Conference, 2003, pp. 647–651.

[34] E. Darvé, J. Solomon, and A. Kiab, *Computing generalized langevin equations and generalized fokkerplanck equations*, PNAS, 106 (2009), pp. 10884–10889.

[35] Elaine P. M. de Sousa, Agma J. M. Traina, Caetano Traina Jr., and Christos Faloutsos, *Evaluating the intrinsic dimension of evolving data streams*, in SAC, 2006, pp. 643–648.

[36] P. Del Moral, *Feynman-Kac Formulae: Genealogical and Interacting Particle Systems with Applications*, Springer New York, 2004.

[37] P. Del Moral, A. Doucet, and A. Jasra, *Sequential Monte Carlo for Bayesian Computation (with discussion)* , in Bayesian Statistics 8, Oxford University Press, 2006.

[38] P. Del Moral, A. Doucet, and A. Jasrau, *Sequential Monte Carlo Samplers*, Journal of the Royal Statistical Society B, 68 (2006), pp. 411–436.

[39] C. Dellago, P.G. Bolhuis, and D. Chandler, *Efficient transition path sampling: Application to lennard-jones cluster rearrangements*, JOURNAL OF CHEMICAL PHYSICS, 108 (1998), pp. 9236 – 9245.

[40] M. Dellnitz, M. Hessel-von Molo, P. Metzner, R. Preis, and C. Schütte, *Graph algorithms for dynamical systems*, in Analysis, Modeling and Simulation of Multiscale Problems, A. Mielke, ed., Springer-Verlag, Heidelberg, 2006, p. 619646.

[41] A.P. Dempster, N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm (with discussion)*, J. Roy. Statist. Soc. Ser. B, 39 (1977), pp. 1–38.

[42] P. Deuflhard, W. Huisinga, A. Fischer, and C. Schütte, *Identification of almost invariant aggregates in reversible nearly uncoupled markov chains*, LINEAR ALGEBRA AND ITS APPLICATIONS, 315 (2000), pp. 39 – 59.

[43] P. Deuflhard and M. Weber, *Robust perron cluster analysis in conformation dynamics*, LINEAR ALGEBRA AND ITS APPLICATIONS, 398 (2005), pp. 161 – 184.

[44] D.L. Donoho and C. Grimes, *Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data*, PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, 100 (2003), pp. 5591 – 5596.

[45] A. Doucet, J. F. G. de Freitas, and N. J. Gordon, eds., *Sequential Monte Carlo Methods in Practice*, Springer, New York, 2001.

[46] A. Doucet, S.J. Godsill, and C. Andrieu, *On Sequential Monte Carlo Sampling Methods for Bayesian Filtering*, Statistics and Computing, 10, pp. 197–208.

[47] W. E and B. Engquist, *The heterogeneous multi-scale methods*, Comm. Math. Sci., 1 (2003), p. 87.

[48] W. E, D. Liu, and E. Vanden-Eijnden, *Analysis of multiscale methods for stochastic differential equations*, Comm. Pure Appl. Math., 58 (2005), pp. 1544–1585.

[49] Weinan E, Weiqing Ren, and Eric Vanden-Eijnden, *Finite temperature string method for the study of rare events.*, J Phys Chem B, 109 (2005), pp. 6688 – 93.

[50] R. Erban, T.A. Frewen, X.A. Wang, T.C. Elston, R. Coifman, B. Nadler, and I.G. Kevrekidis, *Variable-free exploration of stochastic models: A gene regulatory network example*, JOURNAL OF CHEMICAL PHYSICS, 126 (2007), p. 155103.

[51] Christos Faloutsos, Tamara G. Kolda, and Jimeng Sun, *Mining large graphs and streams using matrix and tensor tools*, in SIGMOD Conference, 2007, p. 1174.

[52] I. Fatkullin and E. Vanden-Eijnden, *A computational strategy for multiscale systems with*

applications to lorenz 96 model, JOURNAL OF COMPUTATIONAL PHYSICS, 200 (2004), pp. 605–638.

[53] M. Ferreira and H. Lee, *Multiscale Modeling - A Bayesian Perspective*, Springer Series in Statistics, Spinger, 2007.

[54] M.A.R. Ferreira, M. West, H. Lee, and D. Higdon, *Multiscale and hidden resolution time series models*, Bayesian Analysis, 2 (2006), pp. 294–314.

[55] A. Fischer, S. Waldhausen, I. Horenko, E. Meerbach, and C. Schütte, *Identification of biomolecular conformations from incomplete torsion angle observations by hidden markov models*, JOURNAL OF COMPUTATIONAL CHEMISTRY, 28 (2007), pp. 2453 – 2464.

[56] Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky, *An hdp-hmm for systems with state persistence*, in ICML, 2008, pp. 312–319.

[57] ———, *Nonparametric bayesian learning of switching linear dynamical systems*, in NIPS, 2008, pp. 457–464.

[58] C. Franzke, D. Crommelin, A. Fischer, and A. Majda1, *A hidden markov model perspective on regimes and metastability in atmospheric flows*, J. Climate, 21 (2008), pp. 1740–1757.

[59] B. Ganapathysubramanian and N. Zabaras, *A non-linear dimension reduction methodology for generating data-driven stochastic input models*, Journal of Computational Physics, 227 (2008), pp. 6612–6637.

[60] C.W. Gear, T.J. Kaper, I.G. Kevrekidis, and A. Zagaris, *Projecting to a slow manifold: Singularly perturbed systems and legacy codes*, SIAM JOURNAL ON APPLIED DYNAMICAL SYSTEMS, 4 (2005), pp. 711 – 732.

[61] C.W. Gear and I.G. Kevrekidis, *Projective methods for stiff differential equations: Problems with gaps in their eigenvalue spectrum*, SIAM JOURNAL ON SCIENTIFIC COMPUTING, 24 (2003), pp. 1091 – 1106.

[62] ———, *Telescopic projective methods for parabolic differential equations*, JOURNAL OF COMPUTATIONAL PHYSICS, 187 (2003), pp. 95 – 109.

[63] C.W. Gear, I.G. Kevrekidis, and C. Theodoropoulos, *'Coarse' integration/bifurcation analysis via microscopic simulators: micro-Galerkin methods*, COMPUTERS & CHEMICAL ENGINEERING, 26 (2002), pp. 941 – 963.

[64] Z. Ghahramani, *An Introduction to Hidden Markov Models and Bayesian Networks*, Journal of Pattern Recognition and Artificial Intelligence, 15 (2001), pp. 9–42.

[65] ———, *Unsupervised learning*, in Advanced Lectures on Machine Learning LNAI 3176, O. Bousquet, G. Raetsch, and U. von Luxburg, eds., Springer-Verlag, 2004.

[66] Zoubin Ghahramani and Geoffrey E. Hinton, *Variational learning for switching state-space models*, Neural Computation, 12 (2000), pp. 831–864.

[67] Z. Ghahramani and M.I. Jordan, *Factorial hidden Markov models*, Machine Learning, 29 (1999), pp. 245–273.

[68] D. Givon, R. Kupferman, and A. Stuart, *Extracting macroscopic dynamics: Model problems and algorithms*, Nonlinearity, (2004).

[69] S.J. Godsill, A. Doucet, and M. West, *Monte carlo smoothing for nonlinear time series*, JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION, 99 (2004), pp. 156 – 168.

[70] T. Griffiths and Z. Ghahramani, *Infinite latent feature models and the indian buffet process*, in NIPS, 2006.

[71] E. Hairer, C. Lubich, and G. Wanner, *Geometric numerical integration*, Springer-Verlag, New York, 2002.

[72] J. D. Hamilton, *A new approach to the economic analysis of nonstationary time series and the business cycle*, Econometrica, 57 (1989), p. 357384.

[73] G. Hammond, P. Lichtner, and C. Lu, *Toward petascale computing in geosciences: application to the Hanford 300 Areau*, Journal of Physics: Conference Series, 125 (2008), p. 012051.

[74] P. J. Harrison and C. F. Stevens, *Bayesian forecasting (with discussion)*, Royal Statistical Society B, (1976).

[75] K.A. Heller and Z. Ghahramani, *A Nonparametric Bayesian Approach to Modeling Overlapping Clusters*, in 11th International Conference on AI and Statistics (AISTATS 2007), 2007.

[76] K.A. Heller, S Williamson, and Z. Ghahramani, *Statistical models for partial membership*, in Proceedings of the 25th International Conference on Machine Learning, 2008.

[77] G. Hinton, *Training products of experts by minimizing contrastive divergence*, Neural Computation, 14 (2002).

[78] W.G. Hoover, *Nonequilibrium molecular dynamics*, Nuclear Physics, (1992), pp. 523–536.

[79] I. Horenko, *On simultaneous data-based dimension reduction and hidden phase identification*, JOURNAL OF THE ATMOSPHERIC SCIENCES, 65 (2008), pp. 1941 – 1954.
[80] I. Horenko, R. Klein, S. Dolaptchiev, and C. Schuette, *Automated generation of reduced stochastic weather models i: Simultaneous dimension and model reduction for time series analysis*, MULTISCALE MODELING & SIMULATION, 6 (2007), pp. 1125 – 1145.
[81] I. Horenko, F. Noe, C. Hartmann, and C. Schütte, *Data-based parameter estimation of generalized multidimensional langevin processes*, Physical Review E, 78 (2007), p. 016706.
[82] I. Horenko, J. Schmidt-Ehrenberg, and C. Schütte, *Set-oriented dimension reduction: Localizing principal component analysis via hidden markov models*, COMPUTATIONAL LIFE SCIENCES II, PROCEEDINGS, 4216 (2006), pp. 74 – 85.
[83] I. Horenko and C. Schütte, *Likelihood-based estimation of multidimensional langevin models and its application in biomolecular dynamics*, MULTISCALE MODELING & SIMULATION, 7 (2008), pp. 731 – 773.
[84] H. Hotelling, *Analysis of a complex of statistical variables into principal components*, Journal of Educational Psychology, 24 (1933), pp. 417–441.
[85] Piotr Indyk, Nick Koudas, and S. Muthukrishnan, *Identifying representative trends in massive time series data sets using sketches*, in VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, September 10-14, 2000, Cairo, Egypt, Amr El Abbadi, Michael L. Brodie, Sharma Chakravarthy, Umeshwar Dayal, Nabil Kamel, Gunter Schlageter, and Kyu-Young Whang, eds., Morgan Kaufmann, 2000, pp. 363–372.
[86] R. A. Jacobs, M. I. Jordan, and G. E. Nowlan, S. J. nad Hinton, *Adaptive mixtures of local experts*, Neural Computation, 3 (1991), p. 7987.
[87] M. I. Jordan and R. A. Jacobs, *Hierarchical mixtures of experts and the EM algorithm*, Neural Computation, 6 (1994), p. 181214.
[88] E. Kalnay, *Atmospheric Modeling, Data Assimilation and Predictability*, Cambridge University Press, 2002.
[89] Kothuri Venkata Ravi Kanth, Divyakant Agrawal, Amr El Abbadi, and Ambuj K. Singh, *Dimensionality reduction for similarity searching in dynamic databases*, Computer Vision and Image Understanding, 75 (1999), pp. 59–72.
[90] M.E. Kavousanakis, R. Erban, A.G. Boudouvis, C.W. Gear, and I.G. Kevrekidis, *Projective and coarse projective integration for problems with continuous symmetries*, JOURNAL OF COMPUTATIONAL PHYSICS, 225 (2007), pp. 382 – 407.
[91] Eamonn J. Keogh, Kaushik Chakrabarti, Sharad Mehrotra, and Michael J. Pazzani, *Locally adaptive dimensionality reduction for indexing large time series databases*, in SIGMOD Conference, 2001, pp. 151–162.
[92] I.G. Kevrekidis, C.W. Gear, and G. Hummer, *Equation-free: The computer-aided analysis of complex multiscale systems*, AICHE JOURNAL, 50 (2004), pp. 1346 – 1355.
[93] IG Kevrekidis, CW Gear, JM Hyman, PG Kevrekidis, O Runborg, and K Theodoropoulos, *Equation-free multiscale computation: enabling microscopic simulators to perform system-level tasks*, Communications in Mathematical Sciences, 1 (2003), pp. 715–762.
[94] Flip Korn, H. V. Jagadish, and Christos Faloutsos, *Efficiently supporting ad hoc queries in large datasets of time sequences*, in SIGMOD Conference, 1997, pp. 289–300.
[95] P.S. Koutsourelakis, *Stochastic upscaling in solid mechanics: An excercise in machine learning*, Journal of Computational Physics, 226 (2007), pp. 301–325.
[96] S. Lafon and A.B. Lee, *Diffusion maps and coarse-graining: A unified framework for dimensionality reduction, graph partitioning, and data set parameterization*, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 28 (2006), pp. 1393 – 1403.
[97] A. Laio and M. Parrinello, *Escaping free-energy minima*, PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA, 99 (2002), pp. 12562 – 12566.
[98] S.H. Lam, *Using csp to understand complex chemical-kinetics*, COMBUSTION SCIENCE AND TECHNOLOGY, 89 (1993), pp. 375 – 404.
[99] S.H. Lam and D.A. Goussis, *The csp method for simplifying kinetics*, INTERNATIONAL JOURNAL OF CHEMICAL KINETICS, 26 (1994), pp. 461 – 486.
[100] B. Leimkuhler and S. Reich, *Simulating Hamiltonian Dynamics*, Cambridge University Press, 2004.
[101] T. Li, A. Abdulle, and W. E, *Effectiveness of implicit methods for stiff stochastic differential equations*, Comm. Comput. Phys., 3 (2008), pp. 295–307.
[102] D.O. Lignell, J.H. Chen, and E.S. Richardson, *Terascale direct numerical simulations of*

turbulent combustion fundamental understanding towards predictive models, Journal of Physics: Conference Series, 125 (2008), p. 012031.

[103] M. Liu, F. West, *A dynamic modelling strategy for Bayesian computer model emulation*, Bayesian Analysis, 4 (2009), pp. 393–412.

[104] U. Maas and S.B. Pope, *Simplifying chemical-kinetics - intrinsic low-dimensional manifolds in composition space*, COMBUSTION AND FLAME, 88 (1992), pp. 239 – 264.

[105] I. Mezić, *Spectral properties of dynamical systems, model reduction and decompositions*, Nonlinear Dynamics, 41, pp. 309–325.

[106] R.N. Miller, E.F. Carter, and S.T. Blue, *Data assimilation into nonlinear stochastic models*, Tellus, 51A (1999), pp. 167–194.

[107] B. Nadler, S. Lafon, R.R. Coifman, and I.G. Kevrekidis, *Diffusion maps, spectral clustering and reaction coordinates of dynamical systems*, APPLIED AND COMPUTATIONAL HARMONIC ANALYSIS, 21 (2006), pp. 113 – 127.

[108] E. Ott, B.R. Hunt, I. Szunyogh, A.V. Zimin, E.J. Kostelich, M. Corazza, E. Kalnay, D. Patil, and J.A. Yorke, *A local ensemble Kalman filter for atmospheric data assimilation*, Tellus A, 56 (2004), pp. 415–428.

[109] Spiros Papadimitriou, Anthony Brockwell, and Christos Faloutsos, *Adaptive, unsupervised stream mining*, VLDB J., 13 (2004), pp. 222–239.

[110] Spiros Papadimitriou, Jimeng Sun, and Christos Faloutsos, *Streaming pattern discovery in multiple time-series*, in VLDB, 2005, pp. 697–708.

[111] K. Pearson, *On lines and planes of closest fit to systems of points in space*, The London, Edinburgh and Dublin Philosophical Magazine and Journal of Science, 2 (1901), pp. , 559572.

[112] L.R. Petzold, L.O. Jay, and J. Yen, *Numerical solution of highly oscillatory ordinary differential equations*, Acta Numerica, 7 (1997), pp. 437–483.

[113] Z. Ren and S.B. Pope, *Reduced description of complex dynamics in reactive systems*, Journal of Physical Chemistry A, 111 (2007), pp. 8464–8474.

[114] R. Rico-Martinez, C.W. Gear, and I.G. Kevrekidis, *Coarse projective kmc integration: forward/reverse initial and boundary value problems*, JOURNAL OF COMPUTATIONAL PHYSICS, 196 (2004), pp. 474 – 489.

[115] S.T. Roweis and L.K. Saul, *Nonlinear dimensionality reduction by locally linear embedding*, 290 (2000), p. 2323.

[116] T. Rydén, *Consistent and asymptotically normal parameter estimates for hidden markov models*, Ann. Stat., 22 (1994), p. 18841895.

[117] Yasushi Sakurai, Spiros Papadimitriou, and Christos Faloutsos, *Braid: Stream mining through group lag correlations*, in SIGMOD Conference, 2005, pp. 599–610.

[118] G. Samaey, I.G. Kevrekidis, and D. Roose, *Patch dynamics with buffers for homogenization problems*, JOURNAL OF COMPUTATIONAL PHYSICS, 213 (2006), pp. 264 – 287.

[119] M. Sato and S. Ishii, *On-line em algorithm for the normalized gaussian network*, Neural Computation, 12 (2000), pp. 407–432.

[120] Acharya A. (2006) 195 6287-6311. Sawant, A., *Model reduction via parametrized locally invariant manifolds: Some examples*, Computer Methods in Applied Mechanics and Engineering, 195 (2006), pp. 6287–6311.

[121] B. Schölkopf, A. J. Smola, and K.-R. Mueller, *Kernel principal component analysis*, in LECTURE NOTES IN COMPUTER SCIENCE, SPRINGER VERLAG KG, 1997, pp. 583–588.

[122] S. Vishwanathan S. V.N. Schraudolph, N. N. Gunter, *Fast iterative kernel pca*, in ADVANCES IN NEURAL INFORMATION PROCESSING SYSTEMS, 2007, pp. 1225–1232.

[123] J.B. Shi and J. Malik, *Normalized cuts and image segmentation*, IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 22 (2000), pp. 888 – 905.

[124] R. H. Shumway and D. S. Stoffer, *Dynamic linear models with switching.*, J. Amer. Stat. Assoc., 86 (1991), p. 763769.

[125] N. Srebro and S. Roweis, *Time-varying topic models using dependent dirichlet processes*, Tech. Report UTML TR 2005003,, Department of Computer Science, University of Toronto, 2005.

[126] Jimeng Sun, Spiros Papadimitriou, and Christos Faloutsos, *Distributed pattern discovery in multiple streams*, in PAKDD, 2006, pp. 713–718.

[127] M. Tao, H. Owhadi, and J.E. Marsden, *Non-intrusive and structure preserving multiscale integration of stiff ODEs, SDEs and Hamiltonian systems with hidden slow dynamics via flow averaging.* http://arxiv.org/abs/0908.1241.

[128] M.A. Taylor, J. Edwards, and A. St.Cyr, *Petascale atmospheric models for the communit climate system model: new developments and evaluation of scalable dynamical cores*, Journal of Physics: Conference Series, 125 (2008), p. 012023.
[129] Y Teh, M Jordan, M Beal, and D Blei, *Hierarchical dirichlet processes*, Journal of the American Statistical Association, (2006).
[130] J.B. Tenenbaum, V. de Silva, and J.C. Langford, *A global geometric framework for nonlinear dimensionality reduction*, SCIENCE, 290 (2000), p. 2319.
[131] Wei-Guang Teng, Ming-Syan Chen, and Philip S. Yu, *A regression-based temporal pattern mining scheme for data streams*, in VLDB, 2003, pp. 93–104.
[132] M. Tuckerman, B. J. Berne, and G.J. Martyna, *Reversible multiple time scale molecular dynamics*, J. Chem. Phys., 97 (1992), p. 19902001.
[133] N. Vaswani, *Particle filtering for large dimensional state spaces with multimodal observation likelihoods*, IEEE Trans. Signal Processing, (2008), pp. 4583–4597.
[134] A.F. Voter, F. Montalenti, and T.C. Germann, *Extending the time scale in atomistic simulation of materials*, ANNUAL REVIEW OF MATERIALS RESEARCH, 32 (2002), pp. 321 – 346.
[135] A.F. Voter, F. Montalenti, T.C. Germann, B.P. Uberuaga, and J.A. Sprague, *Accelerated molecular dynamics methods.*, ABSTRACTS OF PAPERS OF THE AMERICAN CHEMICAL SOCIETY, 223 (2002), pp. 238–COMP.
[136] C. Wang, D. Blei, and D. Heckerman, *Continuous time dynamic topic models*, in Uncertainty in Artificial Intelligence [UAI], p. 2008.
[137] M. K. Warmuth and D. Kuzmin, *On-line variance minimization*, in Proceedings of the 19th Annual Conference on Learning Theory (COLT 06), Carnegie Mellon, Pittsburg PA, 2006.
[138] J. Weare, D. Givon, and P. Stinis, *Variance reduction for particle filters of systems with time-scale separation*, IEEE Trans. Signal Proc., 57 (2009), p. 424.
[139] A.S. Weigend and N.A. Gershenfeld, eds., *Time Series Prediction: Forecasting The Future And Understanding The Past*, Santa Fe Institute Studies in the Sciences of Complexity, Westview Press, 1993.
[140] M. West and P. Harrison, *Bayesian Forecasting and Dynamic Models*, New York: Springer-Verlag, 1997.
[141] C.K. Wikle, L.M. Berliner, and N. Cressie, *Hierarchical Bayesian space-time models*, Environmental and Ecological Statistics, 5 (1998), pp. 117–154.
[142] Yong Yao and Johannes Gehrke, *Query processing in sensor networks*, in CIDR, 2003.
[143] Byoung-Kee Yi, Nikolaos Sidiropoulos, Theodore Johnson, H. V. Jagadish, Christos Faloutsos, and Alexandros Biliris, *Online data mining for co-evolving time sequences*, in ICDE, 2000, pp. 13–22.
[144] A. Zagaris, H.G. Kaper, and T.J. Kaper, *Analysis of the computational singular perturbation reduction method for chemical kinetics*, JOURNAL OF NONLINEAR SCIENCE, 14 (2004), pp. 59 – 91.
[145] Yunyue Zhu and Dennis Shasha, *Statstream: Statistical monitoring of thousands of data streams in real time*, in VLDB, 2002, pp. 358–369.
[146] R. Zwanzig, *Noequilibrium Statistical Mechanics*, Oxford University Press, New York, 2001.