Spectral representation and reduced order modeling of the dynamics of stochastic reaction networks via adaptive data partitioning

Khachik Sargsyan, Bert Debusschere, Habib Najm

Sandia National Laboratories, 7011 East Ave., MS 9051, Livermore, CA 94550, USA
{ksargsy, bjdebus, hnnajm}@sandia.gov

Olivier Le Maître

Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur, BP 133, F-91 403, Orsay, France
olm@limsi.fr

Abstract

Dynamical analysis tools are well established for deterministic models. However, for many biochemical phenomena in cells the molecule count is low, leading to stochastic behavior that causes deterministic macroscale reaction models to fail. The main mathematical framework representing these phenomena is based on jump Markov processes that model the underlying *stochastic reaction network*. Conventional dynamical analysis tools do not readily generalize to the stochastic setting due to non-differentiability and absence of explicit state evolution equations. We developed a reduced order methodology for dynamical analysis that relies on the Karhunen-Loève decomposition and polynomial chaos expansions. The methodology relies on adaptive data partitioning to obtain an accurate representation of the stochastic process, especially in the case of multimodal behavior. As a result, a mixture model is obtained that represents the reduced order dynamics of the system. The Schlögl model is used as a prototype bistable process that exhibits time-scale separation and leads to multimodality in the reduced order model.

1 Introduction

The simplest description of chemical reaction processes is based on *rate equations*, i.e. ordinary differential equations (ODEs) for species concentrations. This macroscopic setting

fails when the relevant volume or the species numbers are small because of the increased significance of stochastic noise due to random molecular collisions [23, 45]. *Stochastic reaction networks* (SRNs) account for intrinsic stochastic noise, and provide a general framework for chemical reaction models at the microscopic, molecular level. SRNs are generally governed by the Chemical Master Equation [20] (CME), which is a differential equation governing the time evolution of the Probability Density Function (PDF) of species numbers. The chemical master equation is obtained by modeling a SRN as a jump Markov process [19, 47], i.e. discrete-state, continuous-time stochastic processes with no memory. Since computing direct numerical solutions for CMEs is still challenging (for recent efforts, see [32, 33, 35] and references therein), simulation-based methods become the main analytical tools. In particular, Gillespie's Stochastic Simulation Algorithm [17, 18] (SSA) provides a simulation mechanism for the time-evolution of species numbers at the microscopic scale, thereby effectively sampling the CME solution. This allows determining useful statistical properties of the system by averaging *without* solving the CME itself.

There are well-developed tools [46] for dynamical analysis of features of deterministic systems such as, for instance, attractors, limit-cycles, chaotic behavior, lower-dimensional manifolds. *A priori* analysis methods of stochastic systems have also been extensively discussed [2, 44]. However, such approaches for jump Markov processes are still very much underdeveloped [5], mainly due to the absence of explicit governing state evolution equations. Also, derivative-based characteristics of the system, such as jacobians, eigenmodes and sensitivity coefficients, do not readily generalize to the stochastic context because of the non-differentiability of the time series.

In this paper, we rely on Karhunen-Loève (KL) expansions [15, 25, 31] that represent the underlying stochastic processes in terms of orthonormal random variables, truncated to a reduced order model. This low-order representation is constructed based on the observed statistics of the stochastic process over a given period of time. The larger the number of retained expansion terms, the smaller the energy in the neglected fluctuations

is. With a truncated KL expansion, each realization of a stochastic process corresponds to a finite number of random variables. Note that these random variables are in general uncorrelated, but not independent. Further, they have general, non-standard distributions determined by the data. As a result, it is desirable to represent these random variables with polynomial chaos (PC) expansions. Polynomial chaos representations of random variables enable computationally efficient estimation of system properties. First defined by Wiener [49], PC expansions found applications in porous media transport, thermo-fluid systems, solid mechanics, chemical systems [11–13, 27, 30, 36, 37], and in the general stochastic finite elements context [14, 15].

A distinct feature that sets SRNs apart from the above-mentioned applications is the presence of *intrinsic* noise due to randomness of reaction processes. In order to apply standard orthogonal projection methods to obtain the PC coefficients, one needs to form a one-to-one correspondence between the underlying stochastic sample space and a space of standard random variables that serves as a domain for the PC expansion. In this work, the Rosenblatt transformation [39], which maps any set of jointly distributed random variables to the same number of independent uniform random variables, is implemented to provide such a correspondence.

However, a global PC representation with a finite order and dimensionality does not accurately capture random variables that exhibit strong multimodalities [40]. Adaptive multi-wavelet [26, 28, 29] or PC [48] bases, both relying on stochastic domain decomposition, enable efficient analysis of such processes in the continuous deterministic setting. In this work, we extend the methodology proposed in [40] to obtain an adaptive partitioning that is data-driven, rather than domain-based, and thus is able to capture the structure and modalities of intrinsic stochasticity exhibited by the data. We show that our data partitioning algorithm, which involves a combination of clustering and data range decomposition, leads to a mixture of PC expansions that properly represents multimodal distributions by taking advantage of the underlying data structure.

The paper is organized as follows. In Section 2, the Karhunen-Loève decomposition is introduced as a tool to obtain a reduced order model of a stochastic process. The same section also presents a prototype bistable model - the Schlögl model - as a benchmark process to test the algorithmic developments of this work. Then, Section 3 introduces PC expansions, and provides details of obtaining PC coefficients via an approximate inverse Rosenblatt transformation. Next, the case for decomposition methods is advocated due to the failure of finite global expansions for multimodal distributions. Data partitioning schemes are proposed and discussed in Section 4, where we develop a new hybrid algorithm that involves data clustering and data range bisection. Numerical tests are presented in Section 5 for both a test sample set and one obtained from the KL projection of the Schlögl model. Finally, Section 6 summarizes the results and outlines ongoing and future work.

2 Reduced Order Modeling via Karhunen-Loève Decomposition

In order to simplify the dynamical analysis of a stochastic process, the Karhunen-Loève spectral decomposition is applied. This leads to a reduced order model, which recognizes the temporal scales present in the time series data of the process and effectively reduces the time-dependent process to a countable (in practice, finite) set of uncorrelated random variables.

2.1 Karhunen-Loève Decomposition

Assume a stochastic process $X(t,\theta)$ is defined on a sample space Θ up to time $T < +\infty$, i.e. $\theta \in \Theta$ and $t \in [0,T]$. Denote the mean of this process over sample realizations by $\bar{X}(t) = \langle X(t,\theta) \rangle$, where the angle brackets denote the average over samples. Separating

the mean, one can write

$$X(t,\theta) = \bar{X}(t) + X_0(t,\theta), \tag{1}$$

where $X_0(t,\theta)$ is a *centered* stochastic process, i.e. its mean vanishes for all $t \in [0,T]$.

The covariance function $C(t_1,t_2) = \langle X_0(t_1,\theta)X_0(t_2,\theta)\rangle$ of a centered stochastic process is symmetric, bounded and positive definite. Hence, it can be expanded as a sum [15]

$$C(t_1,t_2) = \sum_{n=1}^{\infty} \lambda_n f_n(t_1) f_n(t_2) \tag{2}$$

with respect to its complete, orthonormal set of eigenfunctions $f_n(t)$ and real positive eigenvalues λ_n (in a descending order) that are defined as solutions of the integral equation

$$\int_0^T C(t_1,t_2) f_n(t_1) dt_1 = \lambda_n f_n(t_2). \tag{3}$$

The underlying stochastic process $X(t,\theta)$ then admits the *KL decomposition* [25,31]

$$X(t,\theta) = \bar{X}(t) + \sum_{n=1}^{\infty} \xi_n(\theta)\sqrt{\lambda_n} f_n(t) \tag{4}$$

in terms of the eigenfunctions $f_n(t)$ and random variables $\xi_n(\theta) : \Theta \to \mathbb{R}$, which satisfy the orthonormality conditions

$$\langle \xi_n(\theta)\rangle = 0 \qquad \text{and} \qquad \langle \xi_n(\theta)\xi_m(\theta)\rangle = \delta_{nm}. \tag{5}$$

Using the orthonormality of the eigenfunctions, one can recover these uncorrelated - but not independent - random coefficients $\xi_n(\theta)$ by projection of the sample trajectories onto the eigenfunction basis:

$$\xi_n(\theta) = \frac{1}{\sqrt{\lambda_n}} \int_0^T X_0(t,\theta) f_n(t) dt. \tag{6}$$

In practice, the sum in Eq. (4) is truncated, leading to a finite KL decomposition in terms of L zero-mean, uncorrelated random variables $\{\xi_i\}_{i=1}^{L}$. Hitherto in this paper,

the argument θ as an element of the sample space is dropped for convenience. The finite KL decomposition essentially reduces the time dependence in the stochastic process to a finite number of generally dependent random variables. More time scales are taken into account as more terms in the decomposition are considered.

2.2 Numerical Tests for a Deterministically Bistable Process

As a benchmark process with a wide time scale separation, the Schlögl model [19, 41] is used. It is a reaction network involving three species X, A and B and two reversible reactions:

$$\begin{aligned} A + 2X &\underset{a_2}{\overset{a_1}{\rightleftarrows}} 3X, \\ B &\underset{a_4}{\overset{a_3}{\rightleftarrows}} X. \end{aligned} \tag{7}$$

The species A and B are assumed to be in large excess compared to X: their numbers will be held fixed. The *propensity functions* a_i for the reactions are:

$$\begin{aligned} a_1 &= k_1 A X(X-1)/2, \\ a_2 &= k_2 X(X-1)(X-2)/6, \\ a_3 &= k_3 B, \\ a_4 &= k_4 X. \end{aligned} \tag{8}$$

Note that, for simplicity, we denote by A, B and X the number of molecules as well as their names. Table 1 presents the nominal parameter set, with inverse-time units for the rates.

Table 1: Nominal parameter set for the Schlögl model

Parameter	Value
$k_1 A$	0.03
k_2	0.0001
$k_3 B$	200
k_4	3.5
A	10^5
B	$2 \cdot 10^5$
$X(0)$	250

The corresponding rate equation for the concentration of X, i.e., the large volume continuum limit, has three stationary points: two stable equilibria and an unstable one in between. As for the discrete state system (7), some realizations fluctuate near the first basin of attraction, and the rest fluctuate near the other one. The nominal parameter values from Table 1 are chosen to lead to a system that behaves in such a way, with approximately equal proportions of realizations ending up near each of the metastable states, see Figure 1.

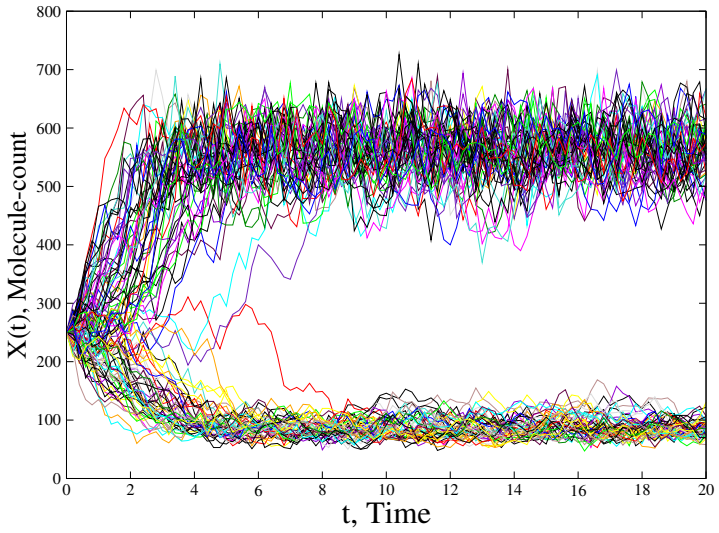

Figure 1: Hundred sample realizations of the Schlögl model with the nominal parameter set of Table 1.

For the Schlögl model, the covariance was discretized with 200 time steps and computed based on 10000 SSA realizations. The KL eigenvalues and corresponding eigenfunctions (called *KL modes*) for this covariance are plotted in Figure 2. All KL modes, except the first one, fluctuate around zero with decreasing amplitudes. The first KL mode seems to saturate to a constant value for this particular time window $[0, 20]$. However, if one extends the time horizon to include values of switching times between the two basins, the first KL mode would also show an oscillatory behavior. Equivalently, if the gap between the first two eigenvalues is smaller (i.e., there is a smaller separation between the two basins of attractions), the switching time would decrease, leading to faster oscillations in

the first KL mode. The large gap between the first and second eigenvalues is explained by the bistability. The first eigenvalue corresponds to the evolution of the system state from the initial state to the vicinity of the upper or lower branches. The truncation in the KL expansion essentially removes the higher-frequency, smaller-amplitude fluctuations. This effect can be seen in the 1-term KL expansion, in Figure 3. It illustrates truncated KL sums for $L = 1, 2, 5, 10, 50, 100$, while Figure 4 shows the corresponding SSA realizations, together with the PDFs of $X(t_f)$ and its KL representations at a fixed time $t_f = 20$. Clearly, as L increases, the KL trajectories converge to those obtained by the SSA, and, consequently, so do the corresponding PDFs.

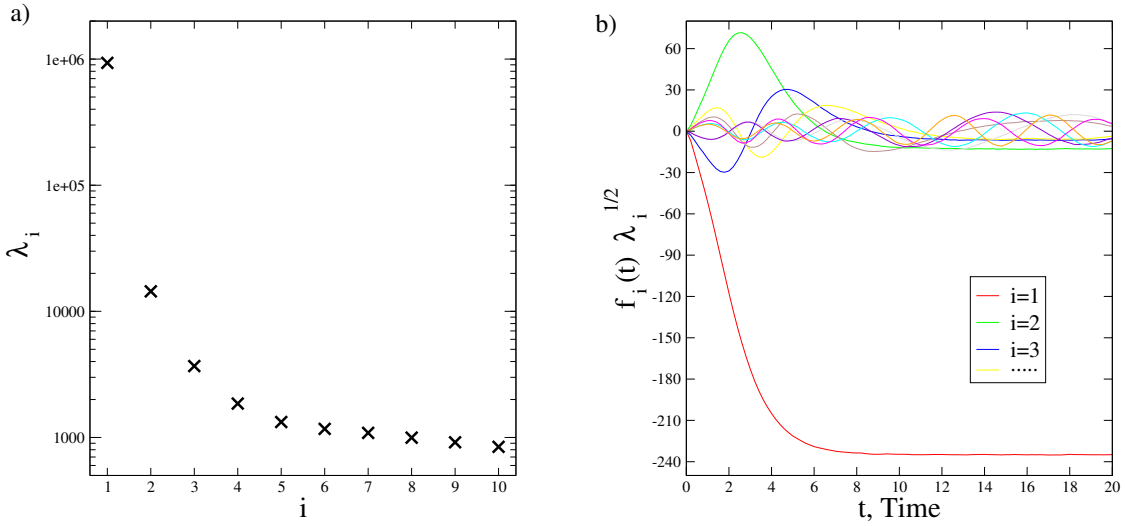

Figure 2: a) First ten KL eigenvalues. b) Rescaled eigenmodes for the Schlögl model with the nominal parameter set of Table 1.

The bimodality of the SSA realizations in the Schlögl model lead to the bimodality of the *KL random variable* ξ_1. Since the first KL mode is much larger than the others (see Figure 2) the sign of the random variable ξ_1 determines whether the particular realization follows the upper or the lower branch. Indeed, the first order KL expansion is $X(t) \approx \bar{X}(t) + \xi_1 \sqrt{\lambda_1} f_1(t)$. The first KL random variable ξ_1 has a bimodal marginal distribution, hence leading to bimodality in realizations. Next, Figure 5 shows a scatter plot of the first two KL random variables ξ_1 and ξ_2. As can be seen, although the variables ξ_1 and

ξ_2 are uncorrelated, there is a strong dependence between them. For example, the range of values of ξ_2 depends on the sign of ξ_1. Consequently, it is important to take this dependence into account in further representations of these variables.

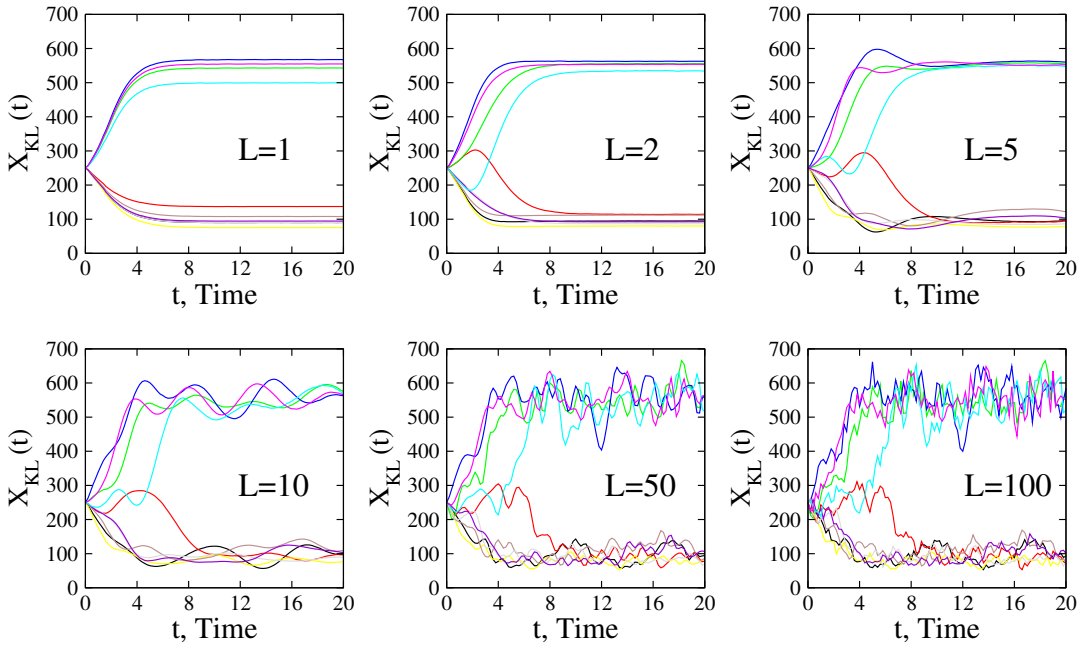

Figure 3: The truncated KL expansion $X_{KL}(t)$ with the number of terms set to $L = 1, 2, 5, 10, 50, 100$. As the number of terms increases, smaller time scales are taken into account.

3 Polynomial Chaos Expansion of the Reduced Order Model

After the original stochastic process is essentially reduced to a random vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_L)$, we use spectral methods to properly represent the latter. In this section, polynomial chaos expansions of random variables are introduced, together with challenges associated with their computations, some of these challenges being specific to SRNs.

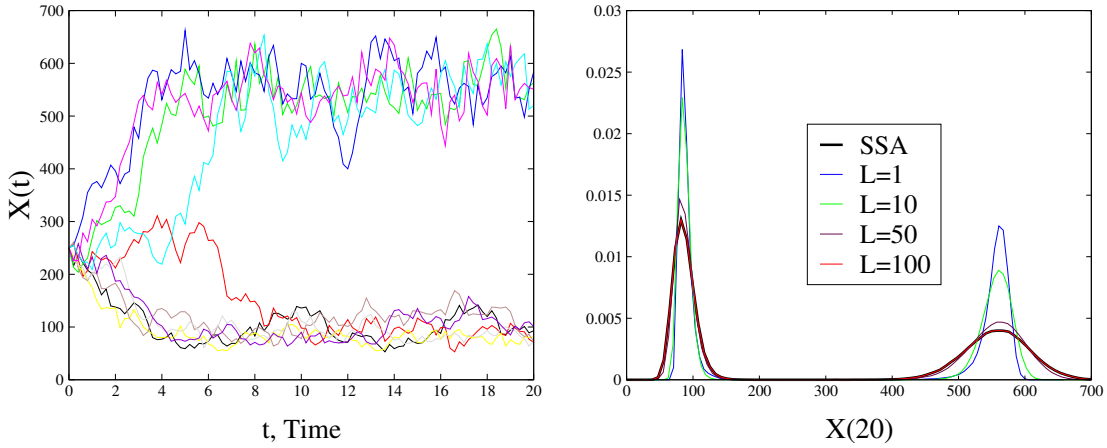

Figure 4: Ten SSA realizations with the nominal parameter set, together with the PDFs of X for SSA realizations and its KL representations (with $L = 1, 10, 50, 100$) at $t_f = 20$ based on 100000 samples.

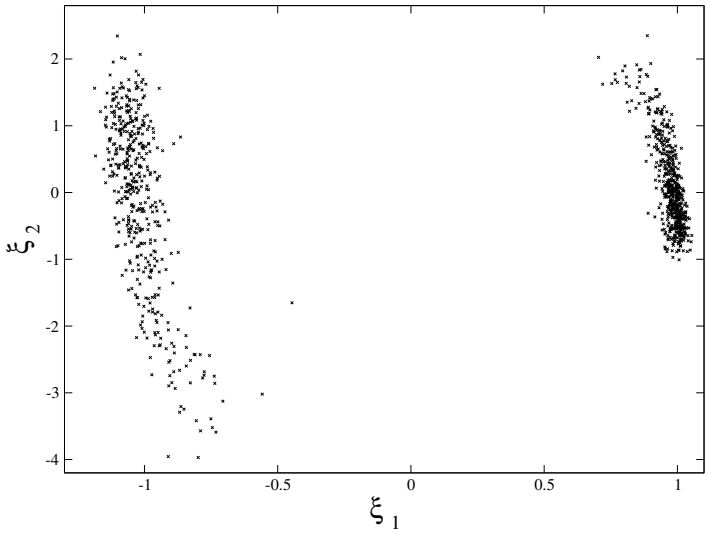

Figure 5: The scatter plot of the first two KL random variables ξ_1, ξ_2. They are uncorrelated, *not* independent, and exhibit strong bimodality in the ξ_1-direction.

3.1 Polynomial Chaos Expansion

Any random variable with finite variance can be expanded in terms of an orthogonal set of functions of specific standard random variables, where the orthogonality is with respect to the density of the latter. These expansions are referred to as polynomial chaos

(PC) expansions [15,49]. Two of the most commonly used expansions for continuous random variables are Gauss-Hermite (Hermite polynomials as functions of standard normal random variables) and Legendre-Uniform (Legendre polynomials as functions of uniform random variables) PC expansions [50]. The major conclusions of this work are independent of the basis choice; we will primarily work with the Gauss-Hermite (GH) PC basis. The Hermite polynomial with a multi-index $\boldsymbol{p} = (p_1, p_2, \ldots, p_L)$ is a multivariate polynomial function of L variables $(\zeta_1, \zeta_2, \ldots, \zeta_L) = \boldsymbol{\zeta}$ defined by

$$\Psi_{\boldsymbol{p}}(\boldsymbol{\zeta}) = \psi_{p_1}(\zeta_1)\psi_{p_2}(\zeta_2)\cdots\psi_{p_L}(\zeta_L), \tag{9}$$

where $\psi_{p_i}(\zeta)$ is the standard one-dimensional Hermite polynomial of degree p_i, for $i = 1, 2, \ldots, L$. The first five Hermite polynomials are

$$\begin{aligned} \psi_0(\zeta) &= 1 \\ \psi_1(\zeta) &= \zeta \\ \psi_2(\zeta) &= \zeta^2 - 1 \\ \psi_3(\zeta) &= \zeta^3 - 3\zeta \\ \psi_4(\zeta) &= \zeta^4 - 6\zeta^2 + 3, \end{aligned} \tag{10}$$

satisfying the recurrence relation [1] for all $m = 1, 2, \ldots$

$$\psi_{m+1}(\zeta) = \zeta\psi_m(\zeta) - m\psi_{m-1}(\zeta). \tag{11}$$

By convention, the sum of all degrees $p_1 + p_2 + \cdots + p_L$ is called the degree of the multidimensional Hermite polynomial (9).

Furthermore, the p-th order, L-dimensional Gauss-Hermite PC expansion of a random

vector $\boldsymbol{\xi} = (\xi_1, \xi_2, \ldots, \xi_L)$ is

$$\boldsymbol{\xi} = \sum_{k=0}^{P} \boldsymbol{c}_k \Psi_k(\boldsymbol{\zeta}) \equiv \boldsymbol{g}(\boldsymbol{\zeta}; \boldsymbol{C}), \tag{12}$$

or

$$\xi_i = \sum_{k=0}^{P} c_{ik} \Psi_k(\boldsymbol{\zeta}), \text{ for } i = 1, 2, \ldots, L, \tag{13}$$

where $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_L)$ is a vector of independent, standard normal random variables and $\boldsymbol{C}$ denotes the matrix of coefficients c_{ik}. In Eq. (12), the scalar subscript in the multidimensional polynomials $\Psi_k(\boldsymbol{\zeta})$ refers to the *graded lexicographic ordering* of the underlying multi-indices [6, 51]. The number of these polynomials that have a degree $\leq p$ is $P+1 = \frac{(p+L)!}{p!L!}$, see [50]. Note that in the one-dimensional case ($L = 1$), P is equal to the order (i.e., the highest degree of the polynomials in that expansion) of the PC expansion p.

The above multivariate Hermite polynomials are orthogonal on $\mathbb{R}^L$ with respect to the probability distribution function of the standard normal random variable. This orthogonality implies Hilbert space projection formulae for the PC coefficients (also referred to as *Galerkin projection* [15]):

$$\boldsymbol{c}_k = \frac{\langle \boldsymbol{\xi} \Psi_k(\boldsymbol{\zeta}) \rangle}{\langle \Psi_k^2(\boldsymbol{\zeta}) \rangle} = \frac{1}{\langle \Psi_k^2(\boldsymbol{\zeta}) \rangle} \int \boldsymbol{\xi} \Psi_k(\boldsymbol{\zeta}) \rho(\boldsymbol{\zeta}) d\boldsymbol{\zeta}, \tag{14}$$

where $\rho(\cdot)$ is the PDF of $\boldsymbol{\zeta}$. The angle brackets here denote the expectation, or the integral, with respect to $\boldsymbol{\zeta}$.

The orthogonality of Hermite polynomials also leads to formulae for the statistical moments of $\boldsymbol{\xi}$ in terms of its PC coefficients. In particular, the mean vector and the covariance matrix can be found from

$$\boldsymbol{\mu} = \langle \boldsymbol{\xi} \rangle = \boldsymbol{c}_0 \quad \text{and} \quad \mathrm{Cov}(\xi_i, \xi_j) = \langle (\xi_i - \mu_i)(\xi_j - \mu_j) \rangle = \sum_{k=1}^{P} c_{ik} c_{jk} \langle \Psi_k^2(\boldsymbol{\zeta}) \rangle. \tag{15}$$

Note that throughout this work, since $\boldsymbol{\xi}$ are found by the KL projection of a centered stochastic process, $\langle \boldsymbol{\xi} \rangle = \mathbf{0}$ and $\mathrm{Cov}(\xi_i, \xi_j) = \delta_{ij}$, see Eq. (5).

Two widely-used approaches for determining the PC coefficients using the Galerkin projection are Intrusive Spectral Projection (ISP) and Nonintrusive Spectral Projection (NISP) [8, 9, 36–38]. The ISP approach is based on a direct application of the Galerkin projection to the governing equations. The governing dynamical equations for X are replaced with equations for the PC coefficients, using the projection (14). However, the absence of a deterministic forward model renders the ISP method inapplicable in the SRN context. In this work we will use the NISP approach, which involves sampling of $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ in order to compute the stochastic projection integral $\langle \boldsymbol{\xi} \Psi_k(\boldsymbol{\zeta}) \rangle$. However, since $\boldsymbol{\xi}$ does not belong to the same stochastic space as $\boldsymbol{\zeta}$, one needs to define a map $\Gamma : \boldsymbol{\zeta} \to \boldsymbol{\xi}$ which preserves the probabilities, i.e. $\Gamma(\boldsymbol{\zeta})$ and $\boldsymbol{\xi}$ have the same distributions. This is achieved by employing the inverse Rosenblatt transformation, as explained in the following section.

3.2 Rosenblatt Transformation and its Approximate Inverse

Given any set of L random variables $\xi_1, \ldots, \xi_L$ with known (or, in an approximate setting, statistically estimated) joint cumulative distribution function (CDF) $F(\xi_1, \ldots, \xi_n)$, one can obtain a set of η_i's that are independent uniform random variables on $[0, 1]$ for all $i = 1, 2, \ldots, L$, using the conditional cumulative distributions

$$\begin{aligned} \eta_1 &= F_1(\xi_1) \\ \eta_2 &= F_{2|1}(\xi_2|\xi_1) \\ \eta_3 &= F_{3|2,1}(\xi_3|\xi_2, \xi_1) \\ &\vdots \\ \eta_L &= F_{L|L-1,\ldots,1}(\xi_L|\xi_{L-1}, \ldots, \xi_1). \end{aligned} \tag{16}$$

This map, denoted by the shorthand notation $\boldsymbol{\eta} = R(\boldsymbol{\xi})$, is called the *Rosenblatt transformation* [39]. Note that the Rosenblatt transformation is not unique: by ordering the ξ_i's in different ways, one can obtain $L!$ different sets of uniform random variables.

In practice the exact joint cumulative distribution $F(\xi_1, \ldots, \xi_L)$ is generally not available. Instead, in this work it is estimated by sampling $\boldsymbol{\xi}$ using a standard Kernel Density Estimator (KDE), see [42, 43]. Assume that we have N samples $\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(N)}$ of a random vector $\boldsymbol{\xi}$. The KDE estimate of its joint probability density function is a sum of N multivariate gaussian functions centered at each data point $\boldsymbol{\xi}^{(n)}$:

$$p_{\boldsymbol{\xi}}(\boldsymbol{x}) = \frac{1}{N\sigma^L(2\pi)^{L/2}} \sum_{n=1}^{N} \exp\left(-\frac{(\boldsymbol{x} - \boldsymbol{\xi}^{(n)})^T(\boldsymbol{x} - \boldsymbol{\xi}^{(n)})}{2\sigma^2}\right) \tag{17}$$

or

$$p_{\xi_1,\ldots,\xi_L}(x_1, \ldots, x_L) = \frac{1}{N\sigma^L(2\pi)^{L/2}} \sum_{n=1}^{N} \exp\left(-\frac{(x_1 - \xi_1^{(n)})^2 + \cdots + (x_L - \xi_L^{(n)})^2}{2\sigma^2}\right), \tag{18}$$

where the *bandwidth* σ should be chosen to balance smoothness and accuracy, see [42, 43] for discussions of the choice of σ.

Now the conditional CDF is KDE-estimated by

$$\begin{aligned}
F_{k|k-1,\ldots,1}(\xi_k|\xi_{k-1}, \ldots, \xi_1) &= \int_{-\infty}^{\xi_k} p_{k|k-1,\ldots,1}(\xi_k^{'}|\xi_{k-1}, \ldots, \xi_1) d\xi_k^{'} \\
&= \int_{-\infty}^{\xi_k} \frac{p_{k,\ldots,1}(\xi_k^{'}, \xi_{k-1}, \ldots, \xi_1)}{p_{k-1,\ldots,1}(\xi_{k-1}, \ldots, \xi_1)} d\xi_k^{'} \\
&\approx \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\xi_k} \frac{\sum_{n=1}^{N} \exp\left(-\frac{(\xi_1 - \xi_1^{(n)})^2 + \cdots + (\xi_k^{'} - \xi_k^{(n)})^2}{2\sigma^2}\right)}{\sum_{n=1}^{N} \exp\left(-\frac{(\xi_1 - \xi_1^{(n)})^2 + \cdots + (\xi_{k-1} - \xi_{k-1}^{(n)})^2}{2\sigma^2}\right)} d\xi_k^{'}
\end{aligned}$$

$$
\begin{aligned}
&= \int_{-\infty}^{\xi_k} \frac{\sum_{n=1}^{N} \exp\left(-\frac{(\xi_1-\xi_1^{(n)})^2+\cdots+(\xi_{k-1}-\xi_{k-1}^{(n)})^2}{2\sigma^2}\right) \times \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\xi_k'-\xi_k^{(n)})^2}{2\sigma^2}\right)}{\sum_{n=1}^{N} \exp\left(-\frac{(\xi_1-\xi_1^{(n)})^2+\cdots+(\xi_{k-1}-\xi_{k-1}^{(n)})^2}{2\sigma^2}\right)} d\xi_k' \\
&= \frac{\sum_{n=1}^{N} \exp\left(-\frac{(\xi_1-\xi_1^{(n)})^2+\cdots+(\xi_{k-1}-\xi_{k-1}^{(n)})^2}{2\sigma^2}\right) \times \Phi\left(\frac{\xi_k-\xi_k^{(n)}}{\sigma}\right)}{\sum_{n=1}^{N} \exp\left(-\frac{(\xi_1-\xi_1^{(n)})^2+\cdots+(\xi_{k-1}-\xi_{k-1}^{(n)})^2}{2\sigma^2}\right)},
\end{aligned} \tag{19}
$$

where $\Phi(z)$ is the CDF of a standard normal random variable. Note that the numerator in (19) differs from the denominator only by an extra factor $\Phi\left(\frac{\xi_k-\xi_k^{(n)}}{\sigma}\right)$ in each summand, allowing an efficient computation scheme.

The above Rosenblatt transformation maps the set of i.i.d. uniform random variables η_i to the random vector $\boldsymbol{\xi}$. However, the projection (14) requires the inverse of the Rosenblatt transformation. Nevertheless, the approximate conditional distributions are monotonic, hence they are guaranteed to have an inverse function. Also note that in order to achieve the map needed for the Galerkin projection (14), an extra function $\Phi(\zeta_i) = \eta_i$ should be applied for $i = 1, \ldots, L$ to switch from a standard normal random variable ζ_i to a uniform random variable η_i. This leads to a KDE-approximated inverse Rosenblatt transformation composed with the $\Phi(\cdot)$ map between ζ_i and η_i, denoted by $\boldsymbol{\xi} = \Gamma(\boldsymbol{\zeta})$. Now the expectation in the Galerkin projection formula (14) is well defined and can be computed in $\boldsymbol{\zeta}$-space, with the help of quadrature integration:

$$
\boldsymbol{c}_k = \frac{\langle \boldsymbol{\xi} \Psi_k(\boldsymbol{\zeta}) \rangle}{\langle \Psi_k^2(\boldsymbol{\zeta}) \rangle} = \frac{1}{\langle \Psi_k^2(\boldsymbol{\zeta}) \rangle} \int \underbrace{\Gamma(\boldsymbol{\zeta})}_{\boldsymbol{\xi}} \Psi_k(\boldsymbol{\zeta}) \rho(\boldsymbol{\zeta}) d\boldsymbol{\zeta}. \tag{20}
$$

3.3 Global Expansion

Putting both KL and PC expansions together, one arrives at the following, *global* representation of the stochastic process $X(t,\theta)$:

$$X(t,\theta) - \bar{X}(t) \approx \sum_{n=1}^{L} \xi_n(\theta)\sqrt{\lambda_n} f_n(t) \approx \sum_{n=1}^{L} \left(\sum_{k=0}^{P} c_{nk} \Psi_k(\boldsymbol{\zeta}) \right) \sqrt{\lambda_n} f_n(t) \tag{21}$$

The representation (21) effectively reduces the time dependence and the intrinsic stochasticity to a matrix of coefficients c_{ik}. It is important to note that the representation in (21) can be interpreted in a *strong* sense, due to the point-by-point correspondence between θ and $\boldsymbol{\xi}$ (Karhunen-Loève), as well as between $\boldsymbol{\xi}$ and $\boldsymbol{\zeta}$ (Rosenblatt).

Data-driven approaches that combine KL projections with PC representation have been used before, e.g. [7, 10]. In [10], a PC representation of the KL-projected random variable is found by maximum likelihood estimation, but *without* taking into account the dependencies between various components of the vector $\boldsymbol{\xi}$. In [7], the Rosenblatt transformation was used to capture these dependencies. However, it has been shown [26, 28, 40, 48], that the global PC representations with fixed finite dimension and order fail to properly represent random variables with multimodal distributions. Figure 6 shows the scatter plot of data samples obtained from the KL projection of the Schlögl model samples and the corresponding samples that are obtained from the subsequent PC expansion, for two parameter regimes: one leading to a unimodal distribution (b), and the other corresponding to the nominal parameters, i.e. leading to a bimodal distribution (a). Even the 10-th order expansion does not properly capture the behavior of a bimodal random variable, while a unimodal random variable is represented well with relatively low order.

Multidomain PC expansions resolve this issue to some extent. The essence of domain decomposition procedures proposed in [28, 29] is splitting domains in $\boldsymbol{\zeta}$-space, when the current representation is not satisfactory according to certain criteria, and then proceeding recursively. In [40], in order to better recognize the modalities, we proposed a different

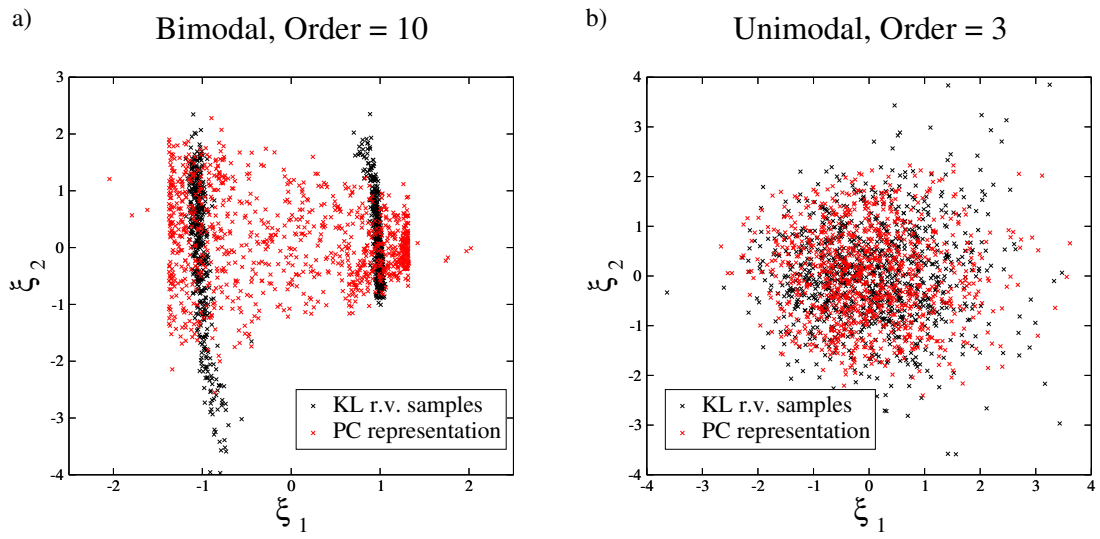

Figure 6: a) Global PC representation is illustrated to fail for the bimodal case, even with the higher order representation. The random vector (ξ_1, ξ_2) is obtained by KL projection of the Schlögl process's time series onto two eigenmodes. b) The unimodal case is obtained by setting the initial value $X(0) = 750$ that leads to a single branch. Both KL and PC expansions are based on 10000 realizations; only 1000 of them are shown for clarity of presentation.

adaptive algorithm that is based on splitting the data domain instead of the $\boldsymbol{\eta}$-domain, and then inferring the PC representations of each partition independently. However, this strategy might be inefficient when dealing with multidimensional and multimodal data structures. In this work, we extend this domain-based data partitioning approach to include data clustering, which finds natural groupings in the data sample set first, effectively reducing the problem to a set of unimodal cases. This results in a faster convergence to data partitions that allow accurate PC representations.

4 Adaptive Data Partitioning Algorithm

In this section, we develop adaptive algorithms that use data partitioning and then proceed recursively until the PC representation is satisfactory according to a specified distance measure. We explore the advantages and shortcomings of two data partitioning strategies: data range bisection and approximate k-center clustering, and introduce a hybrid

approach that uses initial clustering to find natural groupings in the data and then adaptively and recursively partitions the data by splitting corresponding data domains in each direction.

4.1 Data range bisection

Let us consider *range-based bisection* as a data partitioning strategy, i.e. first the smallest and the largest data values in each dimension are found, and then the data itself is partitioned by splitting the range of the data values in half. Here, we will only focus on a brute-force procedure, where all directions are divided into two equal intervals. In [40], we call this strategy 'data domain decomposition'. To avoid confusion with $\boldsymbol{\eta}$-domain decomposition methods [28, 29, 48], we will rather refer to it as range-based bisection. In the current form, the number of new data sets scales exponentially with the dimension L; each subdivision results in 2^L new domains per divided domain. On the other hand, this procedure breaks the data into much smaller and relatively equal-sized sets, leading to a simpler and faster treatment of each subset. However, there is no prior quantitative guarantee on the quality of PC representation for each data partition.

4.1.1 Adaptive algorithm based on Kullback-Leibler divergence

One commonly used measure of the quality of a probabilistic model representation is the *Kullback-Leibler (K-L) divergence or distance* [3, 16] (otherwise called the *relative entropy*) between the sample set PDF $P(\boldsymbol{x})$ and the PDF $Q(\boldsymbol{x})$ of samples obtained from the model:

$$d(P||Q) = \int P(\boldsymbol{x}) \log \frac{P(\boldsymbol{x})}{Q(\boldsymbol{x})} d\boldsymbol{x}. \tag{22}$$

Note that the K-L divergence is not a distance measure in a conventional sense, since it is not symmetric and does not satisfy the triangle inequality. However, it provides an upper bound for the *total variation* distance between two probability measures; $d_{\mathrm{TV}}(P, Q) \leq \sqrt{d(P||Q)/2}$, see [16]. Exact computation of the K-L divergence requires an integration

that is extremely costly in multiple dimensions. Nevertheless, it can be estimated by Monte-Carlo integration in terms of the data samples that are available. Namely,

$$\begin{aligned} d(P||Q) &= \int P(\boldsymbol{x}) \log \frac{P(\boldsymbol{x})}{Q(\boldsymbol{x})} d\boldsymbol{x} \approx \frac{1}{N} \sum_{n=1}^{N} \log \frac{P(\boldsymbol{\xi}^{(n)})}{Q(\boldsymbol{\xi}^{(n)})} \\ &= \frac{1}{N} \left(\sum_{n=1}^{N} \log P(\boldsymbol{\xi}^{(n)}) - \sum_{n=1}^{N} \log Q(\boldsymbol{\xi}^{(n)}) \right), \end{aligned} \tag{23}$$

where $\boldsymbol{\xi}^{(n)}$ for $n = 1, 2, \ldots, N$ are the samples drawn from the distribution $P(\cdot)$, i.e. exactly the data samples that are to be PC-represented. In this form, the K-L divergence allows simple intuitive interpretation: the second sum is the log-probability of having the particular data set $\{\boldsymbol{\xi}^{(n)}\}_{n=1}^{N}$ *given* a model that leads to the PDF $Q(\cdot)$ (in other terms, the *likelihood* of the model), while the first sum is the likelihood *if* the model had the exact same PDF as the original data set (in a sense, a *target* likelihood). The PDFs in (23) are computed by standard KDE techniques [42, 43].

The adaptive PC representation algorithm then proceeds as follows:

0. Obtain N SSA realizations $X(t)$.

1. Perform KL decomposition up to the eigenmode (dimension) L: $X_{KL}(t) = \bar{X}(t) + \sum_{n=1}^{L} \xi_n \sqrt{\lambda_n} f_n(t)$.

   1a. As a result, obtain a set of N data samples of the random vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_L)$ and call it the current data set $S = \{\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(N)}\}$ with Eq. (6).

2. Use the Rosenblatt transformation and quadrature evaluation of the Galerkin projection integrals (20) to find a finite order PC representation for the current data samples: $\xi_i = \sum_{k=0}^{P} c_{ik} \Psi_k(\boldsymbol{\zeta})$, for $i = 1, 2, \ldots, L$.

   2a. Compute the K-L divergence between the data and the PC model with Eq. (23).

3. *If* the number of samples in the current data set exceeds the threshold N_{thr} *and* the K-L divergence is larger than the threshold d_{thr}, partition the current data

set according to data domain decomposition, and recursively return to Step 2 for each new data set. *Else* keep the current PC representation and move to the next untreated data set.

4.1.2 Mixture model and its relation to stochastic domain decomposition

Following the procedure outlined above, the initial data sample set $S = \{\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(N)}\}$ is divided into K_f categories or partitions $S_1, S_2, \ldots, S_{K_f}$, each of them having been treated independently, leading to K_f separate PC representations $\boldsymbol{g}(\boldsymbol{\zeta}; \boldsymbol{C}^{(i)})$, for $i = 1, 2, \ldots, K_f$. Let $|S|$ be the total number of samples in the set S, and $K_h(\cdot)$ be the Gaussian kernel function used in the KDE estimation of the PDFs in (23):

$$K_h(\boldsymbol{z}) = \frac{e^{-\frac{\boldsymbol{z}^T\boldsymbol{z}}{2h^2}}}{h^d(2\pi)^{d/2}}, \tag{24}$$

the KDE-estimated PDF of the data samples is then

$$P(\boldsymbol{y}) = \frac{1}{|S|}\sum_{j=1}^{|S|} K_h(\boldsymbol{y} - \boldsymbol{\xi}^{(j)}). \tag{25}$$

Further, the KDE-estimated PDF of the samples in the k-th data set ($P_k(\boldsymbol{y})$) is K-L approximated by the KDE-estimated PDF of the samples of the k-th PC model $\boldsymbol{g}\left(\boldsymbol{\zeta}; \boldsymbol{C}^{(k)}\right)$, denoted by $Q_k(\boldsymbol{y})$:

$$P_k(\boldsymbol{y}) = \frac{1}{|S_k|}\sum_{j=1}^{|S_k|} K_h(\boldsymbol{y} - \boldsymbol{\xi}_k^{(j)}) \overset{\text{K-L}}{\approx} Q_k(\boldsymbol{y}). \tag{26}$$

Furthermore, the PDF of the full data set can be written as

$$\begin{aligned} P(\boldsymbol{y}) &= \tfrac{1}{|S|}\sum_{j=1}^{|S|} K_h(\boldsymbol{y} - \boldsymbol{\xi}^{(j)}) = \sum_{k=1}^{K_f}\frac{|S_k|}{|S|}\left(\frac{1}{|S_k|}\sum_{j=1}^{|S_k|} K_h(\boldsymbol{y} - \boldsymbol{\xi}_k^{(j)})\right) \\ &= \sum_{k=1}^{K_f}\frac{|S_k|}{|S|}P_k(\boldsymbol{y}) \overset{\text{K-L}}{\approx} \sum_{k=1}^{K_f}\frac{|S_k|}{|S|}Q_k(\boldsymbol{y}) = Q(\boldsymbol{y}), \end{aligned} \tag{27}$$

where $Q(\boldsymbol{y})$ is a mixture PDF of all $Q_k(\boldsymbol{y})$ for $k = 1, \ldots, \mathrm{K}_f$. Hence, we propose the following global model for $\boldsymbol{\xi}$:

$$\boldsymbol{\xi}_{PC} = \boldsymbol{g}(\boldsymbol{\zeta}; \boldsymbol{C}^{(k)}), \text{ with prob. } p_k = |S_k|/|S|, \tag{28}$$

which, in terms of PDFs,

$$Q(\boldsymbol{y}) = \sum_{k=1}^{K_f} \frac{|S_k|}{|S|} Q_k(\boldsymbol{y}) \text{ or } PDF_{\boldsymbol{\xi}_{PC}}(\boldsymbol{y}) = \sum_{k=1}^{K_f} p_k PDF_{\boldsymbol{g}\left(\boldsymbol{\zeta}; \boldsymbol{C}^{(k)}\right)}(\boldsymbol{y}) \tag{29}$$

is simply a mixture model of PDFs of each component. It is worth noting that one can arrive at this mixture representation by using a Bayesian model averaging (BMA) methodology [22], too. We refer to Appendix 1 for rederiving the mixture model (28) with the BMA approach.

Again, the final PC representation is not a single expansion; it is a combination of PC expansions each representing a subset of the data (and each defined on the global stochastic domain) and weighted accordingly to match the PDFs; i.e. the PDF of the final representation is a mixture of the PDFs of samples resulting from each PC expansion. Note that, in principle, the stochastic space decompositions (leading to so-called multidomain expansions) [28, 29, 48] also can be interpreted as mixture models, except that the grouping is based on the underlying sample space elements, rather than the data values themselves. See Appendix 2 for a discussion of the relation between the two approaches.

Note that the range-based bisection is simply one of many possible ways of partitioning the data set. We will also consider two other data-based partitioning strategies: *median-based bisection*, where the data range is divided according to the median value of the data in each dimension, and *size-based bisection*, where the data set is split into two-equal sized subsets consecutively dimension-by-dimension, leading to a final partition of equally-sized (to be precise, each subset will have $N/2^d$ samples) subsets. Figure 7 shows the result of the application of each of the three strategies for an artificially constructed data sample

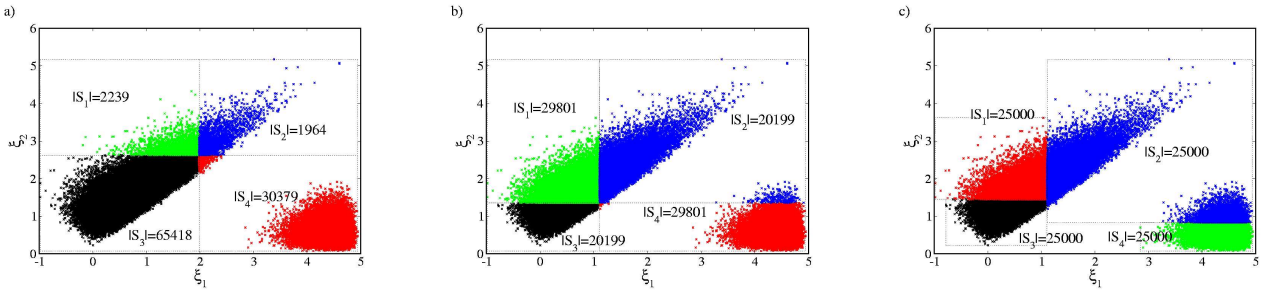

Figure 7: Three different approaches of data partitioning of two-dimensional, bimodal data obtained as a mixture of transformed Gamma distributions, given in (30) : a) range-based bisection, b) median-based bisection, c) size-based bisection.

set of a random vector $\boldsymbol{\xi} = (\xi_1, \xi_2)^T$, where

$$\boldsymbol{\xi} = \begin{cases} T_1\boldsymbol{\chi} & \text{w. prob. } 0.7 \\ T_2\boldsymbol{\chi} + \boldsymbol{s} & \text{w. prob. } 0.3 \end{cases}, \tag{30}$$

with the transformation (scaling and rotation, to be precise) matrices T_1, T_2, as well as the shift vector $\boldsymbol{s}$, defined as

$$T_1 = \begin{pmatrix} 3\cos(\pi/4) & \sin(\pi/4) \\ -3\sin(\pi/4) & \cos(\pi/4) \end{pmatrix}, \; T_2 = \begin{pmatrix} 0 & 1 \\ -0.5 & 0 \end{pmatrix}, \; \boldsymbol{s} = \begin{pmatrix} 5 \\ 0 \end{pmatrix}, \tag{31}$$

and a vector $\boldsymbol{\chi}$ with i.i.d. Gamma$(6, 10)$ distributed components. We will refer to this test distribution as the 'bi-Gamma' distribution.

An evident disadvantage of the domain-based strategies is the inefficiency of their first step; since they do not necessarily detect the two modalities. Therefore in some cases, they can lead to partitions that still have a bimodal structure, making the spectral representation challenging. For the current data set, as can be seen in Figure 7, *all* three strategies lead to new subsets with bimodal structure. For most of the data sets we have considered, the range-based approach works better than the median- and size-based strategies. In particular, for a data set that is obtained from a 2-mode KL projection of

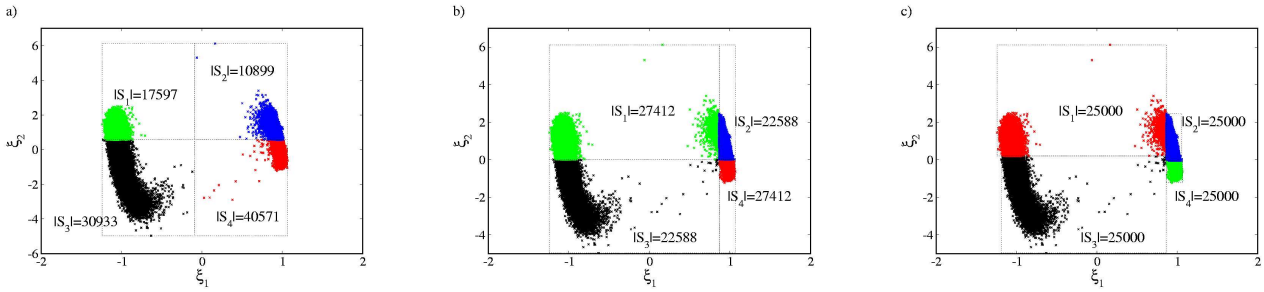

Figure 8: Three different approaches of data partitioning of two-dimensional data obtained from a KL projection of $N = 10^5$ realizations of the Schlögl process: a) range-based bisection, b) median-based bisection, c) size-based bisection.

$N = 10^5$ Schlögl process realizations, the range-based bisection happens to resolve the bimodality, since the data is bimodal along the first dimension, see Figure 8.

Nevertheless, to obtain an algorithm that is more robust and detects the bimodalities independent of the underlying data structure, we turn to clustering approaches in the next section.

4.2 Approximate k-center clustering

In the data clustering literature, the data domain-based approaches from the previous section fall into the category of *monothetic* algorithms, i.e. single features (in our case, the components of the vector $\boldsymbol{\xi}$) of data samples are being compared sequentially, not simultaneously [24]. *Polythetic* algorithms, on the other hand, take all the dimensions into account at the same time, thus generally leading to more sensible groupings of the data. Clustering approaches are well-known to find natural groupings in data sets, according to similarity or distance measures, thus falling in the class of polythetic algorithms.

4.2.1 Formulation

Two of the most commonly used clustering methods are the k-center clustering [4] and the k-means clustering [34]. The k-center clustering problem is posed as follows: given a *predefined* number K, find a partition of the data samples into K clusters $S_1, S_2, \ldots, S_K$

and cluster centroids $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_K$ respectively, that minimizes

$$\max_i \max_{\boldsymbol{\xi} \in S_i} d(\boldsymbol{\xi}, \boldsymbol{s}_i) \tag{32}$$

for some distance measure $d(\boldsymbol{x}, \boldsymbol{y})$. In other words, this is a minimization of the largest cluster, where the size of a cluster is defined as the largest distance from the corresponding centroid. On the other hand, k-means clustering minimizes the average size of a cluster, the latter itself being defined as the average distance from samples in that cluster to the mean (which is the centroid) of the cluster. In some sense, the k-center clustering focuses on the worst scenarios, rather than the average cases, hence it is better at detecting multiple 'islands' of data samples as well as outliers. However, the exact k-center problem is NP-hard. Instead, we will use the *farthest point clustering* greedy algorithm. Its computational complexity scales linearly with the number of samples, and it is proven to lead to a partitioning where the largest cluster is at most twice as large as the exact optimal one [21].

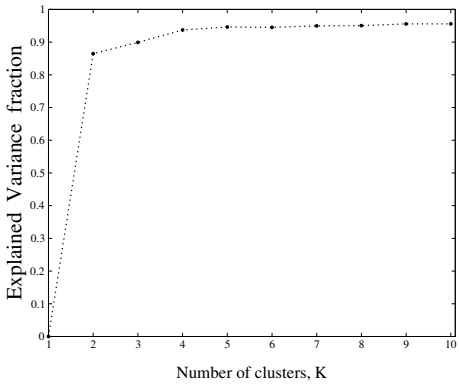

Figure 9: The best explained variance fraction versus the number of clusters for the 'bi-Gamma' data set from (30). For each K, the largest explained variance fraction is shown out of $R = 50$ trials. Clearly, there is a well defined 'elbow' in this graph and any sensible criterion chooses $K = 2$ as the optimal number of clusters for the data set.

It is worth noting that the farthest point clustering algorithm is randomized: it depends on a randomly picked initial centroid. To improve the robustness, we run the clustering algorithm R times (set to $R = 50$, unless noted otherwise; this does not con-

tribute much to the overall computational cost of our algorithms) and pick the outcome that has the largest *explained variance*. The explained variance for a given clustering is the variance of a new sample set that is obtained from the original sample set by replacing each sample with its cluster mean. For example, the explained variance vanishes for just one cluster, and it is equal to the total variance for the maximum possible number of clusters, i.e. if K is equal to the number of samples. With the multiple-run approach, well-separated modalities in data samples are readily detected, and there is less variability in the clustering outcomes for fixed K.

Also note that the k-center clustering requires the number of clusters K to be defined *a priori*. In order to find the optimal number of clusters for a given sample set, the 'elbow' criterion for the explained variance fraction is applied. We run the approximate k-center clustering algorithm R times for each $K = 1, 2, \ldots, K_{\max}$ (typically, $K_{\max} = 10$ is taken), and, for each K, we pick the clustering that leads to the largest explained variance. Figure 9 shows the fraction of this largest explained variance over the total variance versus the number of clusters, for the 'bi-Gamma' data set sampled according to the random vector (30). This graph is generally increasing and concave down. The optimal number of clusters matching the data samples corresponds to the 'elbow' of the graph, i.e. the point where the slope decreases the most with the addition of one cluster. If the x-domain of the graph included real numbers, rather than integers, this 'elbow' would correspond to the largest negative magnitude of the second derivative. Two other reasonable criteria based on the explained variance have been tested, and were found to lead to comparable results for the purposes of this work. Namely, one can set an absolute threshold on the explained variance fraction and choose the smallest K that passes that threshold; or one can simply choose a cluster number K that corresponds to the highest jump in the explained variance fraction.

The other important component of clustering algorithms is the definition of the dis-

tance measure $d(\boldsymbol{x}, \boldsymbol{y})$ between two data points. We will take a weighted L_2 measure

$$d(\boldsymbol{x}, \boldsymbol{y})^2 = \sum_{j=1}^{L} w_j (x_j - y_j)^2, \tag{33}$$

postponing the discussion about the weight choice to Section 5.

4.2.2 Comparison with the data range bisection

Note that k-center clustering partitions the data into a number of clusters that does not depend on the dimensionality of the data, as opposed to data-based bisections. Also, k-center clustering identifies data multimodalities and outliers much better while the data domain bisection strategies blindly divide the data domain. For instance, consider the 'bi-Gamma' random vector (30). While the k-center clustering with the explained variance criterion easily recognizes the two clusters, the 'brute-force' data range bisection only splits the data domains leading to at least one other data set with a bimodal structure, see Figure 10. A drawback of the k-center clustering, however, is revealed when dealing with data sets of relatively simple, unimodal structure. Because of the underlying randomness of the algorithm and the absence of well-defined natural groupings or modalities in the data, the clustering leads to a large variability in the results, as can be seen in Figure 11. In the presence of such a variability, the 'elbow' criterion is useless as the explained variance curve lacks an apparent 'elbow', see Figure 12. In this case, the data is considered to be already sufficiently unimodal, and there is no need to pursue with k-center clustering at all. Hence we modify our explained variance criterion, i.e. we look for an 'elbow' in the explained variance graph *only* if the best explained variance fraction for $K = 10$ is greater than a threshold 0.8. Table 2 outlines the differences between the approximate k-center clustering and the range-based bisection approaches. Note that although the data range bisection leads to a relatively large number of new partitions, each of those partitions are of a much smaller size, allowing faster spectral representations.

In the next section, to combine the advantages of both approaches, we extend the

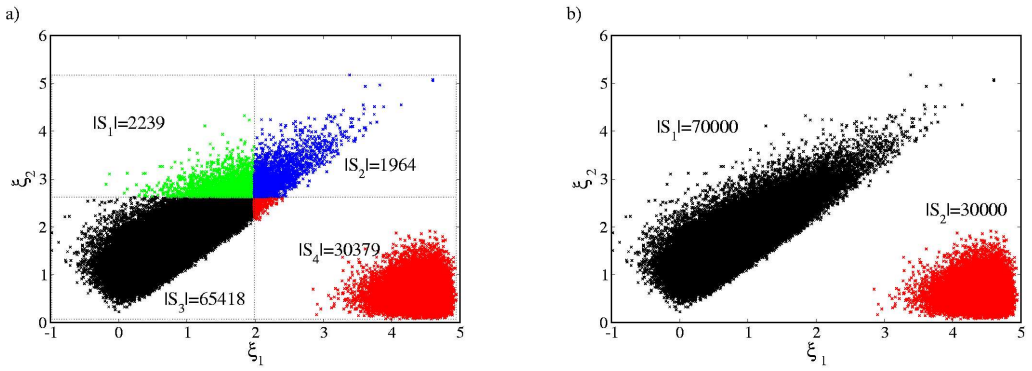

Figure 10: Illustration of a) data range bisection and b) k-center clustering for the 'bi-Gamma' two-dimensional random vector from (30). The 'brute-force' data range bisection blindly splits the data domain into four new ones, without detecting the bimodality. On the other hand, the explained variance criterion, although based on a randomized algorithm, chooses the optimal number of clusters to be equal to two, hence correctly recognizing the underlying data structure.

Approximate k-center clustering	Data range bisection
Polythetic (more effective use of data structure)	Monothetic
Detects multimodalities and outliers	Blind to multimodalities and outliers
No curse of dimensionality	Number of new partitions scales exponentially with dimensions
Dimension-specific weight measure	No weight measure
Non-unique (randomized) partitioning	Unique partitioning
Not effective for unimodal data	Performs well for unimodal data
New subset sizes are of a similar order of magnitude	New subset sizes are extremely reduced

Table 2: Comparison table of features of both partitioning approaches.

data range bisection algorithm of Section 4.1.1 to include an initial k-center clustering step that recognizes the modalities in the data structure and initially partitions the data set according to it.

4.3 Hybrid algorithm

Finding the best number of clusters for a given sample set based on the explained variance criterion leads to an initial natural partitioning that facilitates a better PC representation for each of these partitions or clusters.

The final algorithm is essentially the same as before, with an addition of the initial

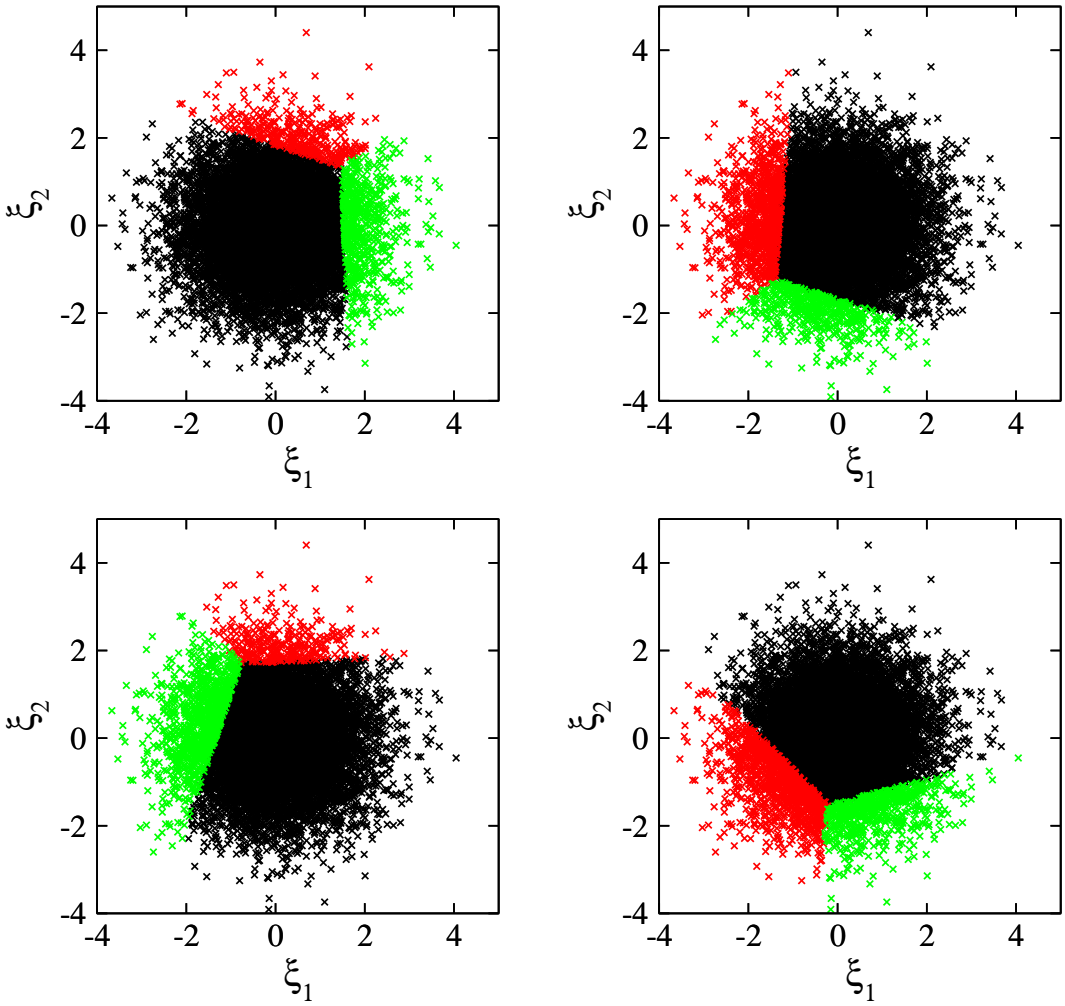

Figure 11: The approximate k-center clustering algorithm is randomized, hence for unimodal data sets the results show huge variability. A two-dimensional gaussian data of $N = 10000$ points is taken. The cluster number is fixed $K = 3$.

clustering step 1b:

0. Obtain N SSA realizations $X(t)$.

1. Perform KL decomposition up to the eigenmode (dimension) L: $X_{KL}(t) = \bar{X}(t) + \sum_{n=1}^{L} \xi_n \sqrt{\lambda_n} f_n(t)$.

   1a. As a result, obtain a set of N data samples of the random vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_L)$ and call it the current data set $S = \{\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(N)}\}$ with Eq. (6).

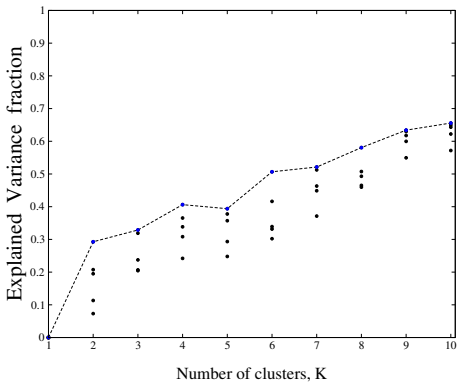

Figure 12: The 'elbow' criterion essentially fails for unimodal data sets. For each fixed cluster number $K = 1, \ldots, 10$, five different runs of the clustering algorithm are shown with the largest explained variance highlighted. There is no well-defined 'elbow' in this graph - a typical feature of all the unimodal data sets.

1b. *If* the explained variance criterion detects modalities, cluster the data into the optimal number of clusters and proceed considering each cluster as a new data set. *Otherwise* proceed to Step 2.

2. Use the Rosenblatt transformation and quadrature evaluation of the Galerkin projection integrals (20) to find a finite order PC representation for the current data samples: $\xi_i = \sum_{k=0}^{P} c_{ik}\Psi_k(\boldsymbol{\zeta}),$ for $i = 1, 2, \ldots, L$.

2a. Compute the K-L divergence between the data and the PC model with Eq. (23).

3. *If* the number of samples in the current data set exceeds the threshold N_{thr} *and* the K-L divergence is larger than the threshold d_{thr}, partition the current data set according to data domain decomposition, and recursively return to Step 2 for each new data set. *Else* keep the current PC representation and move to the next untreated data set.

Eventually, the data will be hierarchically partitioned into a number of sets, say, K_f. Then, the final PC representation can be recovered as a mixture model, see Eqs. (28) and (29). It is worth noting that the hybrid algorithm suffers from the curse of dimensionality, starting from the domain decomposition step. However, it offers huge improvements

compared to plain domain decomposition approaches essentially by replacing the initial unnecessary domain splitting steps with a clustering technique.

5 Numerical Tests

In this section, we present the results of the adaptive hybrid algorithm for two data sets: the data set of samples from the test 'bi-Gamma' random vector (30) and the data set of samples obtained from KL projection of the Schlögl model.

5.1 Data obtained from the 'bi-Gamma' test distribution

The adaptive hybrid algorithm of Section 4.3 leads to a partitioning of the data set obtained from the 'bi-Gamma' test distribution (30). Figure 13 illustrates the resulting partitioning and the samples from the final PC mixture model representation. As can be seen, the first level of partitioning simply recognizes the bimodal structure and separates two unimodal clusters, while the subsequent refinement levels perform data range bisection. As the threshold parameters tighten, the highest refinement level increases. The threshold parameters taken for the illustration in Figure 13 lead to a refinement level of 3. Next, Figure 14 shows the convergence, in terms of the K-L distance, of the mixture PC representation to the original data set, for the hybrid algorithm that includes the clustering step, as well as for the data bisection approaches of Section 4.1.2. The convergence comparison in terms of the K-L distance is shown with respect to refinement level, as well as in terms of the number of subsets generated, K_f and the CPU time. Clearly, the hybrid methodology is more robust and converges faster than the plain domain-based, monothetic approaches.

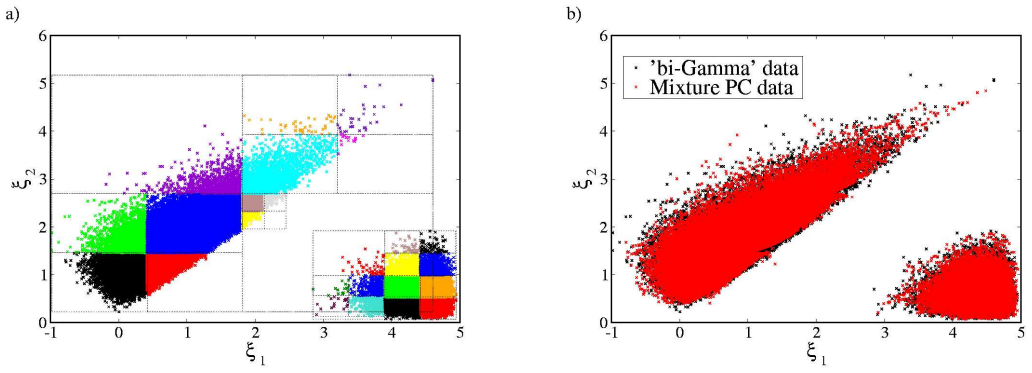

Figure 13: a) The data partitions found by the hybrid algorithm for a data set of $N = 10^5$ samples of the 'bi-Gamma' random vector (30). The threshold parameters are set to $N_{\text{thr}} = 1000$ and $d_{\text{thr}} = 0.003$. b) The scatter plot of the original data set and samples obtained from the mixture PC representation.

5.2 Data obtained from the Schlögl model

If the data is taken from the Karhunen-Loève decomposition of a stochastic process, each dimension has an associated eigenvalue that indicates the importance of that particular dimension. This is particularly important for multimodal systems, since they exhibit large gaps in the eigenvalue spectrum. Therefore for KL-projected data samples, to weigh each dimension according to its importance, an eigenvalue-weighted L_2 measure is taken in (33) with $w_j = \lambda_j$.

As shown in Figure 16, data median bisection and size bisection do not converge well, since they blindly split the data without detecting the bimodality. Even though all three data-based bisection algorithms split the data set to significantly smaller partitions, the unresolved bimodalities for the median- and size-based bisection algorithms require much higher CPU times for the spectral representation because of the higher computational cost of the inverse Rosenblatt transformation specifically for bimodal data sets. Data range bisection works as well as the hybrid algorithm by chance, because of the structure of the bimodality of this particular data set (well-separated by the first dimension).

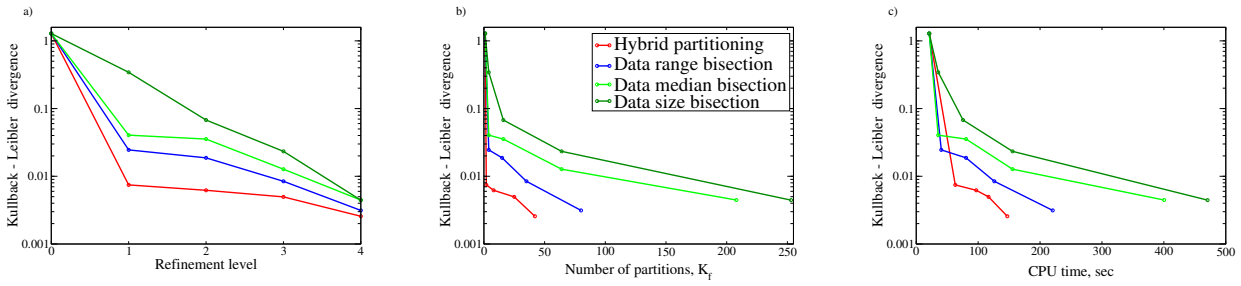

Figure 14: a) Convergence of the mixture PC representation as the refinement level increases. The zeroth refinement level corresponds to the global representation, while the first level is simply the clustering for the hybrid partitioning. The third refinement level corresponds to the illustration from Figure 13. b) c) Convergence with respect to the final number of generated subsets, K_f, and the computational cost of the algorithm. The legend in the middle figure corresponds to all three graphs.

Finally, Figure 17 illustrates the 5-mode KL-truncated sum

$$X_{KL}(t) = \bar{X}(t) + \sum_{n=1}^{5} \xi_n \sqrt{\lambda_n} f_n(t) \tag{34}$$

of the underlying Schlögl process as well as the process, recovered from the third order mixture PC representation of the KL-projected variables

$$X_{KLPC}(t) = \bar{X}(t) + \sum_{n=1}^{5} (\xi_{PC})_n \sqrt{\lambda_n} f_n(t). \tag{35}$$

Since the PDF of the vector $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_5)$ is approximated well by a mixture of PC model PDFs, the resulting reduced order model in Figure 17a) is also approximated well in a distributional sense, i.e. at each time point $\tilde{t}$ the PDFs of $X_{KL}(\tilde{t})$ and $X_{KLPC}(\tilde{t})$ are close to each other. Thus, the stochastic process $X(t)$, first reduced to $X_{KL}(t)$ (described by a random vector $\boldsymbol{\xi}$) by the KL projection, is further reduced to $X_{KLPC}(t)$ (described by a set of deterministic matrices $\{\boldsymbol{C}^{(k)}\}_{k=1}^{K_f}$, one for each partition of the data samples of $\boldsymbol{\xi}$) by our mixture PC representation while preserving the skeleton of the dynamics of the original process for further analysis of the system. Although the PC representation does not add any new information than what is already carried by SSA realizations, it

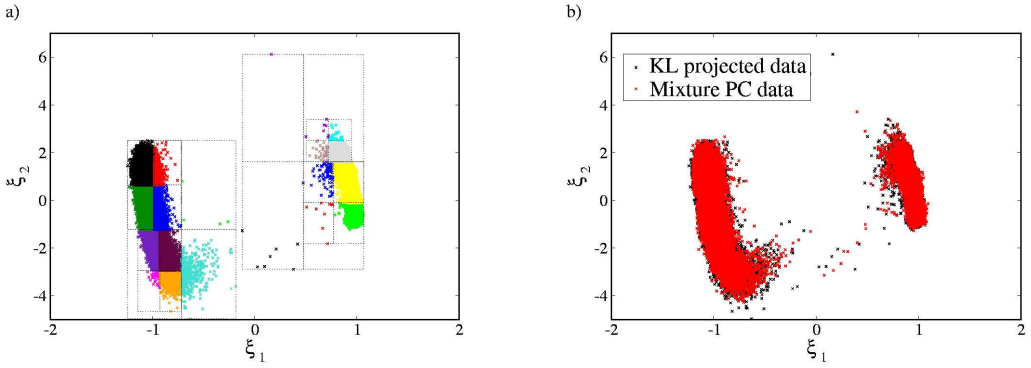

Figure 15: a) The data partitions for the first two KL variables obtained from a KL projection of $N = 10^5$ realizations of the Schlögl process. The threshold parameters are set to $N_{\text{thr}} = 1000$ and $d_{\text{thr}} = 0.01$. b) The scatter plot of the original data set and samples obtained from the mixture PC representation.

allows a finite and manageable representation of the multidimensional stochastic system state. This representation accounts for intrinsic noise in the system and can be extended to propagate parametric uncertainties through this system.

6 Conclusions

In this work, we introduced a methodology for dynamical analysis of stochastic processes relying upon reduced order modeling as well as stochastic spectral expansions for representing intrinsic noise. The Karhunen-Loève expansion has been employed to represent the time-dependent stochastic process in terms of a finite number of random variables. The latter then have been represented using Gauss-Hermite PC expansions. However, since finite global PC expansions do not represent multimodal distributions well, we introduced an adaptive procedure that recognizes the natural groupings in the data and PC-represents each of these partitions separately. In order to perform the orthogonal projection required for obtaining the PC coefficients, the Rosenblatt transformation has been used, mapping the sample space responsible for the intrinsic noise, to a sample space of a standard random variable. In order to reduce the problem to unimodal distributions,

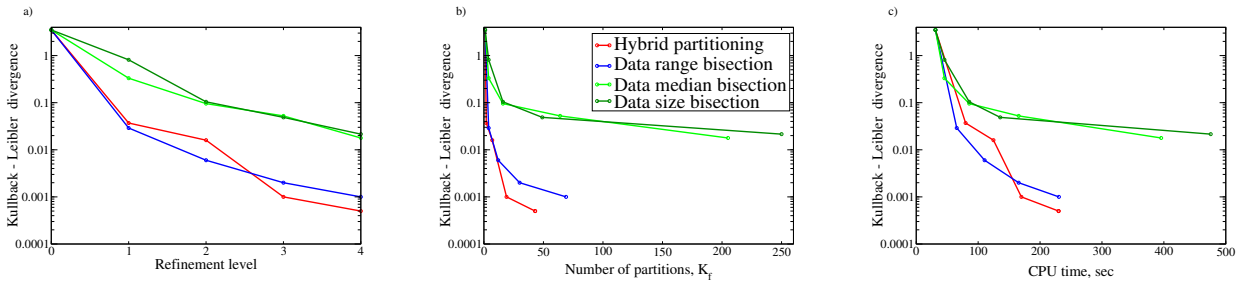

Figure 16: a) Convergence of the mixture PC representation as the refinement level increases. The zeroth refinement level corresponds to the global representation, while the first level is simply the clustering for the hybrid partitioning. The third refinement level corresponds to the illustrations from Figure 15. b) c) Convergence with respect to the final number of generated subsets, K_f, and the computational cost of the algorithm. The legend in the middle figure corresponds to all three graphs.

an approximate version of the k-center clustering algorithm has been used, detecting multimodalities present in the KL-projected data structure, followed by a data partitioning for further refinement. The end result is a combination of PC representations leading to a distribution that is a mixture of GH PC densities. In a sense, this is a step forward from the well-known Gaussian mixture models, allowing not just a simple Gaussian variable, but a polynomial function of it, in each mixture. The overall algorithm presents an efficient approach to obtain a spectral expansion for a stochastic process, especially when the data drawn from the process is multimodal. This adaptive, hybrid data-partitioning algorithm has been demonstrated to work well on the Schlögl model - a benchmark model that exhibits bimodality.

Some noteworthy extensions of the current work should be pointed out. It is worth including parametric uncertainty in the dynamical analysis, essentially extending previous work [40] with the methods proposed here. Also, the approximate inverse Rosenblatt transformation and subsequent quadrature integration for the spectral projection implemented here becomes inefficient with the increase of dimensionality. Methods to address these issues are the subject of ongoing work.

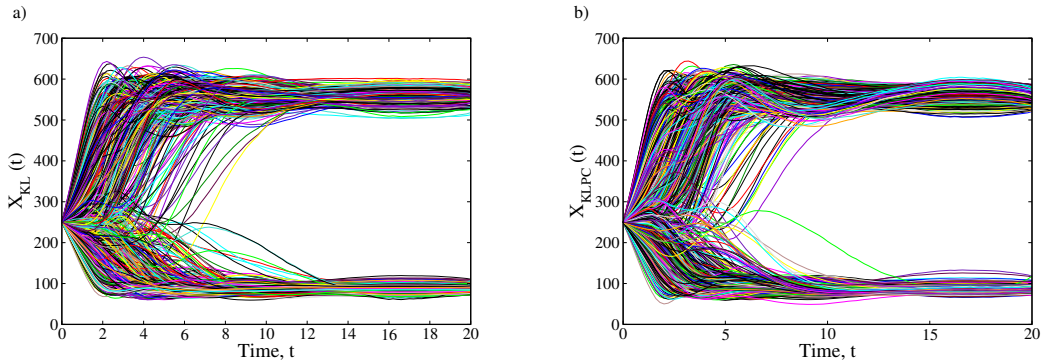

Figure 17: a) The 5-mode KL truncated sum for the Schlögl process. b) The process' final representation obtained from the mixture PC expansions of the underlying five-dimensional KL random vector. Both expansions are obtained with $N = 10^5$ realizations of the Schlögl process with nominal parameters. For illustration purposes, only every hundredth realization is shown.

Acknowledgments

This work was supported by the U.S. Department of Energy Office of Science through the Applied Mathematics program in the Office of Advanced Scientific Computing Research (ASCR) under contract 07-012783 with Sandia National Laboratories. Sandia National Laboratories is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the U.S. Department of Energy under contract No. DE-AC04-94-AL85000.

Appendix 1: Relation to Bayesian Model Averaging

One can derive the mixture model (29) based simply on concepts of Bayesian model averaging (BMA) theory [22]. Indeed, each data partition $S_k = \{\boldsymbol{\xi}_k^{(1)}, \ldots, \boldsymbol{\xi}_k^{(|S_k|)}\}$ is used to inform the construction of a PC model $M_k \equiv \boldsymbol{g}(\boldsymbol{\zeta}; \boldsymbol{C}^{(k)})$ involving the parameter matrix $\boldsymbol{C}^{(k)}$. BMA incorporates all the information from single partitions into a global model in

a consistent manner. The basic statement of BMA reads as

$$p(\boldsymbol{y}|S) = \sum_{k=1}^{K_f} p(\boldsymbol{y}|M_k, S)P(M_k|S). \tag{36}$$

Now, since the data sets are independent,

$$P(M_k|S) = \sum_{i=1}^{K_f} P(M_k|S_i)p(S_i) \tag{37}$$

Further, since, by the construction of the algorithm, the k-th PC model M_k is informed only by the k-th data set S_k, we have

$$P(M_k|S_i) = P(M_k|S_k)\delta_{ki}, \tag{38}$$

or, combining (37) and (38)

$$P(M_k|S) = P(M_k|S_k)p(S_k) \text{ for all } k = 1, \ldots, K_f. \tag{39}$$

Note further that $P(M_k|S_k) = 1$, since $\sum_{i=1}^{K_f} P(M_i|S_k) = 1$ and $P(M_i|S_k)|_{i\neq k} = 0$, per Eq.(38). Thus,

$$P(M_k|S) = p(S_k) \text{ for all } k = 1, \ldots, K_f. \tag{40}$$

Next, since the evidence for the data is obtained by the available samples, we have

$$p_k \equiv p(S_k) = \frac{|S_k|}{|S|} \tag{41}$$

leading to the sought mixture model (compare it to Eq. (29))

$$p(\boldsymbol{y}|S) = \sum_{k=1}^{K_f} p_k p(\boldsymbol{y}|M_k, S). \tag{42}$$

Appendix 2: Relation to Multidomain Expansions

Given a non-overlapping partition $\cup_{i=1}^{K_f} Q_i = [0,1]^L$ of the stochastic space $[0,1]^L$, multidomain expansions can be written as

$$\xi_n = \sum_{i=1}^{K_f} \sum_{k=0}^{P} c_{nk}^{(i)} \tilde{\Psi}_k^{(i)}(\boldsymbol{\eta}), \text{ for } n = 1, \ldots, L, \tag{43}$$

with $\boldsymbol{\eta} \in [0,1]^L$, and where the basis functions $\tilde{\Psi}_k^{(i)}(\cdot)$ vanish outside their support Q_i, usually rectangular, in a form

$$Q_i = [a_i^{(1)}, b_i^{(1)}] \times \cdots \times [a_i^{(L)}, b_i^{(L)}]. \tag{44}$$

The basis functions are either rescaled (straightforwardly, since Q_i's are rectangular) versions of standard PC bases [48] (possibly composed with, say, standard Gaussian inverse CDF $\Phi^{-1}(\cdot)$ for GH PC) or multiwavelets [26, 28, 29]. In the cited works above, the partition $\{Q_i\}_{i=1}^{K_f}$ is obtained from adaptive domain splitting procedures, similar to the one proposed in the beginning of section 4. However, these domain splits are conceptually different from the current work, since they are performed in the stochastic $\boldsymbol{\eta}$-domain $[0,1]^L$. On the other hand, note that the equality (43), if interpreted in a distributional sense, leads to a mixture model (28) with weights $p_i = \text{Area}(Q_i)$. The reverse, however, is not true in general. A mixture model is a distributional equality by definition, and although it can be written in the strong form (43), the stochastic space partition $\cup_{i=1}^{K_f} Q_i = [0,1]^L$ that corresponds to the data partitioning $\cup_{i=1}^{K_f} S_i = S$ is not uniquely defined. In [40], we chose a rectangular partitioning of the $\boldsymbol{\eta}$-space that readily maps to a data domain partition, in order to incorporate the parametric dependence in a straightforward, 'physical' manner. For instance, the first refinement level of the data range bisection would in fact correspond to a rectangular partition of the $\boldsymbol{\eta}$-domain, illustrated in Figure 18. However, if one partitions the data according to, say, clustering approaches discussed in this work,

the data domains may overlap (see Figure 19) and would not map straightforwardly to a $\boldsymbol{\eta}$-domain partition, where the overlaps are impossible *by definition*. Nevertheless, in the present context, there is no reason to specify a particular $\boldsymbol{\eta}$-domain partition, as the mixture model representation is understood in a distributional sense, without point-to-point correspondence between $\boldsymbol{\eta}$ samples and the data ($\boldsymbol{\xi}$ samples).

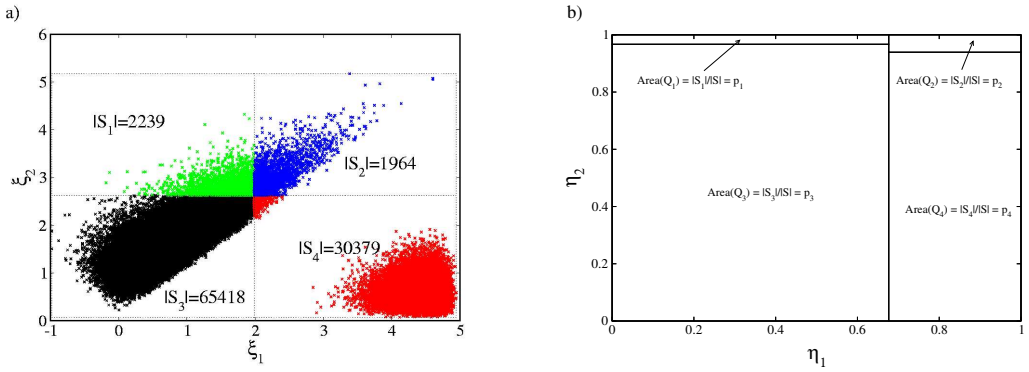

Figure 18: Illustration of the range-based bisection and the corresponding $\boldsymbol{\eta}$-domain decomposition for the 'bi-Gamma' data set (30). Note that, because of the structure of this particular data set, the upper two partitions are of considerably smaller sizes than the lower ones, hence the corresponding rectangles in the $\boldsymbol{\eta}$-domain are smaller, too.

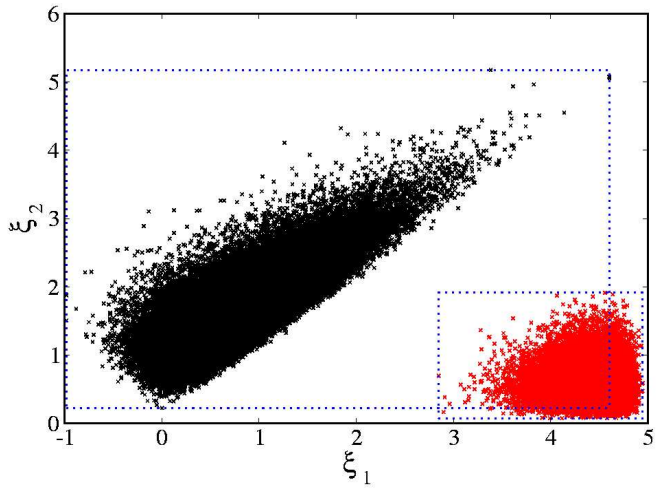

Figure 19: Two well defined clusters in the data set might lead to an overlap between data domains of each cluster, hence not allowing a sensible map to the $\boldsymbol{\eta}$-domain partitions.

To reinforce the relation between data-based bisection and $\boldsymbol{\eta}$-domain decomposition,

consider also the size-based bisection introduced in Section 4.1.2. This case would correspond to the 'brute-force' $\boldsymbol{\eta}$-domain decomposition by bisection [40], as illustrated in Figure 20.

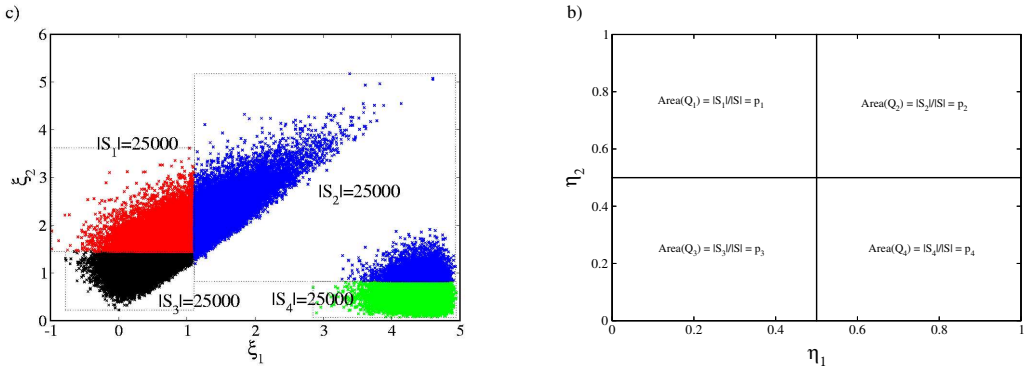

Figure 20: Illustration of the size-based bisection and the corresponding $\boldsymbol{\eta}$-domain decomposition for the 'bi-Gamma' data set (30).

Once again, in this work, a distributional representation is sought, hence we have a freedom of choosing any partitioning in the data domain, while not having to specify an $\boldsymbol{\eta}$-domain decomposition. Although any rectangular $\boldsymbol{\eta}$-domain decomposition with $\text{Area}(Q_i) = p_i$ will lead to a multidomain expansion that is *in distribution* equivalent to the mixture model under consideration, there is no rectangular partitioning in the data domain that would effectively split the two modalities.

References

[1] M. Abramowitz and I. Stegun. *Handbook of Mathematical Functions*. Dover, New York, 1970.

[2] L. Arnold. *Random Dynamical Systems*. Springer, Berlin, 1998.

[3] M. Arnst and R. Ghanem. Probabilistic equivalence and stochastic model reduction in multiscale analysis. *Comput. Methods Appl. Mech. and Engrg.*, 197:3584–3592, 2002.

[4] M. Bern, D. Eppstein, and D. S. Hochbaum. *Approximation algorithms for geometric problems*. PWS Publishing Company, Boston, 1997.

[5] J. Bower and H. Bolouri, editors. *Computational Modeling of Genetic and Biochemical Networks*. The MIT Press, 2001.

[6] D. A. Cox, J. Little, and D. O'Shea. *Ideals, Varieties and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra*. Springer, 1997.

[7] S. Das, R. Ghanem, and J. Spall. Asymptotic sampling distribution for polynomial chaos representation from data: a maximum entropy and Fisher information matrix approach. *SIAM J. Sci. Comp.*, 30(3):2207–2234, 2008.

[8] B. Debusschere, H. Najm, A. Matta, O. Knio, R. Ghanem, and O. Le Maître. Protein labeling reactions in electrochemical microchannel flow: Numerical simulation and uncertainty propagation. *Physics of Fluids*, 15(8):2238–2250, 2003.

[9] B. Debusschere, H. Najm, A. Matta, T. Shu, O. Knio, R. Ghanem, and O. Le Maître. Uncertainty Quantification in a Reacting Electrochemical Microchannel Flow Model. In *Proc. 5^{th} Int. Conf. on Modeling and Simulation of Microsystems*, pages 384–387, San Juan, Puerto Rico, 2002.

[10] C. Desceliers, R. Ghanem, and C. Soize. Maximum likelihood estimation of stochastic chaos representations from experimental data. *International Journal for Numerical Methods in Engineering*, 66:978–1001, 2006.

[11] R. Ghanem. Probabilistic characterization of transport in heterogeneous media. *Comput. Methods Appl. Mech. Engrg.*, 158:199–220, 1998.

[12] R. Ghanem. Ingredients for a general purpose stochastic finite element formulation. *Comput. Methods Appl. Mech. Engrg.*, 168:19–34, 1999.

[13] R. Ghanem. Stochastic finite elements for heterogeneous media with multiple random non-gaussian properties. *ASCE J. Eng. Mech.*, 125:26–40, 1999.

[14] R. Ghanem and J. RedHorse. Propagation of probabilistic uncertainty in complex physical systems using a stochastic finite element approach. *Physica D*, 133(1/4):137 – 144, SEP 1999.

[15] R. Ghanem and P. Spanos. *Stochastic Finite Elements: A Spectral Approach*. Springer Verlag, New York, 1991.

[16] A. Gibbs and F. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70:419–436, 2002.

[17] D. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.*, 22:403–434, 1976.

[18] D. Gillespie. Exact Stochastic Simulation of Coupled Chemical Reactions. *Journal of Physical Chemistry*, 81(25):2340–2361, 1977.

[19] D. Gillespie. *Markov Processes: An Introduction for Physical Scientists*. Academic Press, San Diego, CA, 1992.

[20] D. Gillespie. A rigorous derivation of the chemical master equation. *Phys. A*, 188:404–425, 1992.

[21] T. Gonzalez. Clustering to minimize the maximum intercluster distance. *Theor. Comp. Science*, 38:293–306, 1985.

[22] J. Hoeting, D. Madigan, A. Raftery, and C. Volinsky. Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14(4):382–417, 1999.

[23] W. Horsthemke and R. Lefever. *Noise-Induced Transitions*. Springer-Verlag, Berlin, 1984.

[24] A. K. Jain, M. N. Murty, and P. J. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[25] K. Karhunen. Zur spektraltheorie stochastischer prozesse. *Ann. Acad. Sci. Fennicae*, 37, 1946.

[26] O. Le Maître, R. Ghanem, O. Knio, and H. Najm. Uncertainty propagation using Wiener-Haar expansions. *J. Comput. Phys.*, 197(1):28–57, 2004.

[27] O. Le Maître, O. Knio, H. Najm, and R. Ghanem. A stochastic projection method for fluid flow I. Basic formulation. *J. Comput. Phys.*, 173:481–511, 2001.

[28] O. Le Maître, H. Najm, R. Ghanem, and O. Knio. Multi-resolution analysis of Wiener-type uncertainty propagation schemes. *J. Comput. Phys.*, 197:502–531, 2004.

[29] O. Le Maître, H. Najm, P. Pébay, R. Ghanem, and O. Knio. Multi-resolution-analysis scheme for uncertainty quantification in chemical systems. *SIAM J. Sci. Comput.*, 29(2):864–889, 2007.

[30] O. Le Maître, M. Reagan, H. Najm, R. Ghanem, and O. Knio. A stochastic projection method for fluid flow II. Random process. *J. Comput. Phys.*, 181:9–44, 2002.

[31] M. Loève. *Probability Theory*. Van Nostrand, Princeton, NJ, 1955.

[32] S. MacNamara, A. Bersani, K. Burrage, and R. Sidje. Stochastic chemical kinetics and the total quasi-steady-state assumption: application to the stochastic simulation algorithm and chemical master equation. *Journal of Chemical Physics*, 129(9):095105–1–13, 2008.

[33] S. Macnamara, K. Burrage, and R. Sidje. Multiscale modeling of chemical kinetics via the master equation. *Multiscale Modeling and Simulation*, 6(4):1146–1168, 2008.

[34] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability*, 1:281–297, 1967.

[35] B. Munsky and M. Khammash. The finite state projection approach for the analysis of stochastic noise in gene networks. *IEEE Transactions on Automatic Control*, 53(1):201–14, 2008.

[36] M. Reagan, H. Najm, B. Debusschere, O. Le Maître, O. Knio, and R. Ghanem. Spectral stochastic uncertainty quantification in chemical systems. *Combustion Theory & Modeling*, 8:607–632, 2004.

[37] M. Reagan, H. Najm, R. Ghanem, and O. Knio. Uncertainty quantification in reacting flow simulations through non-intrusive spectral projection. *Combustion and Flame*, 132:545–555, 2003.

[38] M. Reagan, H. Najm, P. Pébay, O. Knio, and R. Ghanem. Quantifying uncertainty in chemical systems modeling. *Int. J. Chem. Kin.*, 37(6):368–382, 2005.

[39] M. Rosenblatt. Remarks on a multivariate transformation. *Annals of Mathematical Statistics*, 23(3):470 – 472, 1952.

[40] K. Sargsyan, B. Debusschere, H. Najm, and Y. Marzouk. Bayesian inference of spectral expansions for predictability assessment in stochastic reaction networks. *J. Comput. Theor. Nanosci.*, 6(10), 2009.

[41] F. Schlögl. On thermodynamics near a steady state. *Z. Phys.*, 248:446–458, 1971.

[42] D. Scott. *Multivariate Density Estimation. Theory, Practice and Visualization*. Wiley, New York, 1992.

[43] B. Silverman. *Density Estimation*. Chapman and Hall, London, 1986.

[44] E. Simiu. *Chaotic Transitions in Deterministic and Stochastic Dynamical Systems*. Princeton University Press, Princeton, NJ, 2002.

[45] R. Srivastava, L. You, J. Summers, and J. Yin. Stochastic vs. deterministic modeling of intracellular viral kinetics. *Journal of Theoretical Biology*, 218(3):309 – 321, 7 2002.

[46] S. Strogatz. *Nonlinear Dynamics and Chaos*. Addison-Wesley, Reading, MA, 1994.

[47] N. van Kampen. *Stochastic Processes in Physics and Chemistry*. Elsevier Science, Amsterdam, 1992.

[48] X. Wan and G. E. Karniadakis. An adaptive multi-element generalized polynomial chaos method for stochastic differential equations. *J. Comput. Phys.*, 209:617–642, 2005.

[49] N. Wiener. The homogeneous chaos. *Am. J. Math.*, 60:897–936, 1938.

[50] D. Xiu and G. Karniadakis. The Wiener-Askey polynomial chaos for stochastic differential equations. *SIAM J. Sci. Comp.*, 24(2):619–644, 2002.

[51] C. Yang, R. Duraiswami, N. A. Gumerov, and L. Davis. Improved fast Gauss transform and efficient kernel density estimation. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 464, Washington, DC, 2003. IEEE Computer Society.