SAND2012-9446P

Project report for Stochastic Dynamical Systems: Analysis of Dynamics and Predictability

Khachik Sargsyan, Cosmin Safta, Bert Debusschere, Habib Najm
Sandia National Laboratories, 7011 East Ave., MS 9051,
Livermore, CA 94550, USA
Email: {ksargsy, csafta, bjdebus, hnnajm}@sandia.gov

October 29, 2012

Contents

Chapter 1

Introduction

Stochastic noise is prevalent in a wide variety of systems, especially when the system behavior is affected or controlled by phenomena on a nanoscale, molecular level, where thermal noise introduces intrinsic variability in molecular interactions. Common examples of such phenomena are chemical or biochemical reactions between small numbers of molecules, found in gene regulation, cell signaling, or interfacial electrochemistry. Given the relevance of these processes in applications ranging from bioremediation and bioenergy (bacterial behavior), biomedicine (immune system signaling), to electrical storage (electrodes), effective tools are needed for the simulation and analysis of such Stochastic Dynamical Systems (SDSs).

The intrinsic noise in these SDSs requires them to be modeled as continuous time, discrete state Markov processes, with state probabilities evolving according to the *master equation* [35]. As the models for such systems are often inferred from noisy observations of a subset of the system state, at sparse instances in time, these SDSs generally also have large levels of parametric and model uncertainties [16, 31], further complicating their simulation and analysis. The project discussed in this report specifically focused on the development of methods for the simulation and analysis of Stochastic Reaction Networks (SRNs), which model the behavior of a set of (bio)chemical reactions between small numbers of molecules. The evolution of SRN state probabilities follows a specific form of the master equation, generally referred to as the Chemical Master Equation (CME) [13].

While the CME provides a full description of the associated stochastic process, its solution is computationally challenging except for the simplest of cases. For this reason, most studies have relied on sampled time trajectories of the system state to examine the system behavior. Such trajectories can be obtained exactly using the Stochastic Simulation Algorithm (SSA) [12], commonly referred to as the Gillespie algorithm. For each sampled trajectory, the algorithm advances the system state one reaction event at the time, by drawing the specific reactions and their time from appropriately constructed distributions. Based on this fundamental simulation method, many advances have been made over the years in both the simulation [1–3, 6, 10, 11, 17, 26] and analysis [15, 19, 25, 27, 28] of stochastic reaction networks.

The first part of this report discusses the development of a reponse surface approach that is particularly robust to noise for uncertainty quantification and sensitivity analysis in high-dimensional SRNs. The approach is applied to the analysis of the transition to competence in *Bacillus subtilis*, based on SSA samples of system trajectories.

Beyond sampling, new approaches have recently become available to solve the CME directly, made possible by advances in algorithms for finite state projection [22, 23], matrix exponentiation [20, 21, 33], spectral representations [4, 8], and sparse quadrature [18]. The combination of

these algorithms has led to dramatic speedups in the solution of the CME, enabling the study of small SRNs by solving their CME directly. As the CME solution offers a full probability distribution of the system states, rather than just samples, it is able to better capture low probability events, which are often missed by sampling-based methods unless a prohibitively large number of samples is used. While such a direct solution is still quite challenging, especially for systems with many molecular species, this approach has been shown for small systems to be competitive with and complementary to the SSA sampling based approaches (see e.g. [21]), and is poised to rapidly find a firm foothold in the community.

The second part of this report discusses the development of a mixed discrete – continuum formulation (discrete CME coupled with continuum Fokker Planck approximation), which speeds up the simulation of SRNs by using the accurate, but costly, discrete CME only where needed. This approach is complementary to, and can potentially be used along with the other approaches discussed in the previous paragraph for accelerating the solution of the CME.

Chapter 2

Response Surface Approach for Uncertainty Quantification in High-Dimensional Stochastic Systems

Stochastic systems with uncertain input parameters carry two major sources of uncertainty - the parametric uncertainty and the inherent stochasticity of the system. For the characterization and quantification of the uncertainties in stochastic systems we employ Polynomial Chaos (PC) spectral expansions. While the PC machinery is well-developed for deterministic systems, it is not as established for stochastic systems. This work is a step towards filling the gap - a PC methodology is implemented for the uncertainty quantification (UQ) in stochastic systems with emphasis on systems with high-dimensional input and a strongly non-linear input-output relationship.

In high-dimensional models, construction of a PC expansion, or, equivalently, a response surface representation of the output dependence on inputs, is challenged by the availability of simulation data. Indeed, PC construction relies on a number of forward model simulations for various input parameter settings. In a high-dimensional input space these input parameter samples generally cover the space only sparsely, particularly for computationally costly stochastic systems. This brings yet another source of uncertainty into the picture - the uncertainty associated with lack-of-knowledge. Bayesian methods are generally well fit to handle such problems and quantify uncertainties associated with the lack-of-knowledge in a consistent manner. We apply a Bayesian inference approach to infer PC expansions that serve as a surrogate model or response surface for the relationship between uncertain inputs and stochastic outputs. Such uncertain PC surrogate models can then replace the full model in simulation-intensive studies, such as global sensitivity analysis, optimization or calibration. While inputs of such PC models correspond to input uncertainties, the PC spectral coefficients themselves are also uncertain due to both the intrinsic stochasticity and the lack-of-knowledge.

Furthermore, a potential drawback of polynomial-based response surface constructions is their inherent smoothness assumption. In practice, strongly non-linear or even discontinuous input-output forward models are very common. In this regard, we develop a data-driven, classification-based approach that leads to a mixture of PC expansions, each of them acting on an input subdomain with a sufficiently smooth output response.

We have demonstrated the methodology to the stochastic reaction network for noise-induced competence transition in *Bacillus subtilis*, a gram-positive soil bacterium with relevance to bioenergy, bioremediation and enchanced oil recovery [5, 14, 24, 30]. The developed techniques lead to a multi-parameter spectral representation of the competence probability with respect to rate constants of in-

dividual reactions of the network. Global sensitivity studies made possible by this approach revealed the most important reactions in terms of their contribution to output uncertainty. This work has recently been published in the *IEEE Transactions on Computational Biology and Bioinformatics* [29] (see Appendix for the final electronic version).

Chapter 3

Mixed Discrete – Continuum Modeling of Stochastic Reaction Networks

This chapter covers the development of a novel, mixed discrete – continuum approach for simulating stochastic reaction networks (SRNs). The motivation behind the approach is to use the discrete CME formulation only for those parts of the system state space where discrete effects matter the most and the CME is essential for accuracy (*e.g.* in the areas where some or all species are present in small numbers of molecules). In the areas where species are present in large numbers of molecules, a continuum Fokker Planck equation (FPE) approximation is applied instead. In areas where only some species are present in sufficiently large numbers of molecules, a hybrid CME/FPE approach is applied, using the discrete representation only for the species present in small numbers of molecules.

As the discretized FPE formulation can use grid sizes that are significantly larger than 1, especially in areas where the probability mass gradients are small, the mixed CME – FPE approach can simulate the full system with a much smaller system of equations than if the CME was used everywhere, thereby reducing the computational cost considerably. Moreover, since the actual system of equations to be solved has the same format as the system of equations that results from the pure CME approach, approaches to accelerate the CME solution (such as finite state projection) should be applicable to the system of equations resulting from the mixed CME – FPE approach.

The next sections first outline the CME, FPE, and hybrid CME – FPE formulations, followed by a discussion of their implementation and testing on SRNs with two species.

3.1 Discrete CME Formulation

Consider an SRN with a d-dimensional state space: $\boldsymbol{N} = (N_1, N_2, \ldots, N_d) \in \mathbb{N}_0^d$, where $\mathbb{N}_0$ denotes the set of all non-negative integers. Denote the (jump, propensity) pairs by $\{(\boldsymbol{\nu}_i, w_r(\boldsymbol{N}))\}_{r=1}^R$, where R is the number of all reactions. More specifically, $\boldsymbol{N} \to \boldsymbol{N} + \boldsymbol{\nu}_r$ with probability $w_r(\boldsymbol{N})$ per unit time.

The process is fully specified by the probabilities $p(\boldsymbol{n}; t) = P\{\boldsymbol{N}(t) = \boldsymbol{n}\}$ that evolve according to the chemical master equation (CME):

$$\frac{d}{dt}p(\boldsymbol{n}; t) = \sum_{r=1}^{R} p(\boldsymbol{n} - \boldsymbol{\nu}_r; t) w_r(\boldsymbol{n} - \boldsymbol{\nu}_r) - \sum_{r=1}^{R} p(\boldsymbol{n}; t) w_r(\boldsymbol{n}). \tag{3.1}$$

As a function of $\boldsymbol{n}$, $p(\boldsymbol{n}; t)$ is called a probability mass function (PMF). Let us order, starting from

index 0 for convenience. the PMFs of all possible states into one vector $\boldsymbol{p}(t)$. Denote the map from a state to its index by $i(\cdot) : \boldsymbol{N} \to \mathbb{N}$. Then, the CME can be written in a matrix form

$$\frac{d\boldsymbol{p}(t)}{dt} = A\boldsymbol{p}(t), \tag{3.2}$$

where

$$A_{kj} = \begin{cases} \delta_{j,i(\boldsymbol{n})} \displaystyle\sum_{r:i(\boldsymbol{n}-\boldsymbol{\nu}_r)=k} w_r(\boldsymbol{n}) & \text{when } j \neq k \\ -\delta_{j,i(\boldsymbol{n})} \displaystyle\sum_{r} w_r(\boldsymbol{n}) & \text{when } j = k \end{cases} \tag{3.3}$$

3.2 Continuous FPE Formulation

In the continuum representation, the discrete states are replaced by the continuous concentrations $\boldsymbol{x} = \boldsymbol{n}/V$, where V is the system volume parameter. The propensities, the jump vectors and the probabilities are also rescaled

$$\bar{w}_r(\boldsymbol{x}) = \frac{1}{V} w_r(\boldsymbol{n}), \qquad \boldsymbol{\xi}_r = \frac{\boldsymbol{\nu}_r}{V}, \qquad \bar{p}(\boldsymbol{x}; t) = V p(\boldsymbol{n}; t). \tag{3.4}$$

For the purposes of this text, the value of V is not relevant. Therefore, we will choose volume units so that $V = 1$. Also, for notational simplicity, we will drop the bars. The context and the arguments (discrete $\boldsymbol{n}, \boldsymbol{\nu}$ or continuous $\boldsymbol{x}, \boldsymbol{\xi}$) will resolve the ambiguity.

The Fokker-Planck equation (FPE) corresponding to the SRN introduced above is

$$\frac{\partial p(\boldsymbol{x}; t)}{\partial t} = -\nabla \cdot (\boldsymbol{f}(\boldsymbol{x}) p(\boldsymbol{x}; t)) + \frac{1}{2} \nabla^2 (\boldsymbol{g}(\boldsymbol{x}) p(\boldsymbol{x}; t)), \tag{3.5}$$

where the *drift* vector and the *diffusion* matrix are defined by

$$\boldsymbol{f}(\boldsymbol{x}) = \sum_{r=1}^{R} \boldsymbol{\xi}_r w_r(\boldsymbol{x}) \qquad \text{and} \qquad \boldsymbol{g}(\boldsymbol{x}) = \sum_{r=1}^{R} \boldsymbol{\xi}_r \boldsymbol{\xi}_r^T w_r(\boldsymbol{x}). \tag{3.6}$$

It is convenient to write FPE in a conservative form

$$\frac{\partial p(\boldsymbol{x}; t)}{\partial t} = \sum_{r=1}^{R} \nabla \cdot \boldsymbol{F}^{(r)}, \tag{3.7}$$

where

$$\boldsymbol{F}^{(r)} = \boldsymbol{\xi}_r \left(-w_r(\boldsymbol{x}) p(\boldsymbol{x}; t) + \frac{1}{2} \boldsymbol{\xi}_r \cdot \nabla \left(w_r(\boldsymbol{x}) p(\boldsymbol{x}; t) \right) \right). \tag{3.8}$$

3.3 Hybrid Approach: CME-FPE

Physical processes often operate at different scales in terms of the number of molecules involved. For some processes, involving small numbers of molecules, the discrete CME approach is necessary

to capture the reaction dynamics while for others, involving species present in large numbers of molecules, the continuous FPE formulation is sufficient to resolve all dynamical scales.

In this section we introduce a formulation [34] that uses the discrete formulation for some species, and the continuum approximation for others. Starting from eq. (3.1), we will split the original $\boldsymbol{n}$ into the ensemble of species molecule numbers that require the discrete formulation, henceforth referred to with the same notation $\boldsymbol{n}$, and $\boldsymbol{m}$ the species are approximated by the continuum formulation

$$\frac{d}{dt}p(\boldsymbol{n},\boldsymbol{m};t) = \sum_{r=1}^{R} p(\boldsymbol{n}-\boldsymbol{\nu}_r,\boldsymbol{m}-\boldsymbol{\mu}_r;t)w_r(\boldsymbol{n}-\boldsymbol{\nu}_r,\boldsymbol{m}-\boldsymbol{\mu}_r) - \sum_{r=1}^{R} p(\boldsymbol{n},\boldsymbol{m};t)w_r(\boldsymbol{n},\boldsymbol{m}). \quad (3.9)$$

In the next step we add and subtract $p(\boldsymbol{n},\boldsymbol{m}-\boldsymbol{\mu}_r;t)w_r(\boldsymbol{n},\boldsymbol{m}-\boldsymbol{\mu}_r)$ to the *rhs* of eq. (3.9)

$$\begin{aligned}\frac{d}{dt}p(\boldsymbol{n},\boldsymbol{m};t) &= \sum_{r=1}^{R} p(\boldsymbol{n}-\boldsymbol{\nu}_r,\boldsymbol{m}-\boldsymbol{\mu}_r;t)w_r(\boldsymbol{n}-\boldsymbol{\nu}_r,\boldsymbol{m}-\boldsymbol{\mu}_r) \\ &\quad -p(\boldsymbol{n},\boldsymbol{m}-\boldsymbol{\mu}_r;t)w_r(\boldsymbol{n},\boldsymbol{m}-\boldsymbol{\mu}_r) \\ &\quad +\sum_{r=1}^{R} p(\boldsymbol{n},\boldsymbol{m}-\boldsymbol{\mu}_r;t)w_r(\boldsymbol{n},\boldsymbol{m}-\boldsymbol{\mu}_r) - p(\boldsymbol{n},\boldsymbol{m};t)w_r(\boldsymbol{n},\boldsymbol{m}).\end{aligned} \quad (3.10)$$

The first two terms in eq. (3.10) correspod to changes that will be approximated by the FPE formulation while the second part will follow the discrete CME formulation. These can be cast in matrix form as

$$\frac{d}{dt}p(\boldsymbol{n},\boldsymbol{m};t) = \sum_{r=1}^{R}(A^{(r)}_{\mathrm{FPE}} + A^{(r)}_{\mathrm{CME}})\cdot p(\boldsymbol{n},\boldsymbol{m};t) \quad (3.11)$$

3.4 Numerical Approach

This section outlines the numerical implementation of this approach for systems with two species and a maximum jump size of 1 in each reaction.

3.4.1 CME

Figure 3.1 shows a schematic of the CME domain and its interfaces with the adjacent domains. The locations of the states in the CME domains are shown with black circles and correpond to integer species counts in the x and y directions respectively. The first couple of layers of states in the adjacent domains are shown with open circles, blue for the hybrid CME-FPE domains, and red for the FPE domain. A buffer layer of states, shown with filled grey circles is necessary to provide boundary conditions for the time advancement of the solution inside the CME domain. The width of this buffer layer is equal to the largest jump size in the kinetic model. For the models in the present study this is equal to one.

Linear interpolations are used to fill in the probability values in the buffer layer. For the buffer layer states in the hybrid CME-FPE region, shown with grey-filled blue circles in Fig. 3.1, one-dimensional interpolation stencils are used. Bi-linear two-dimensional interpolation stencils are used for the buffer layer states in the FPE region, shown with grey-filled red circles in Fig. 3.1.

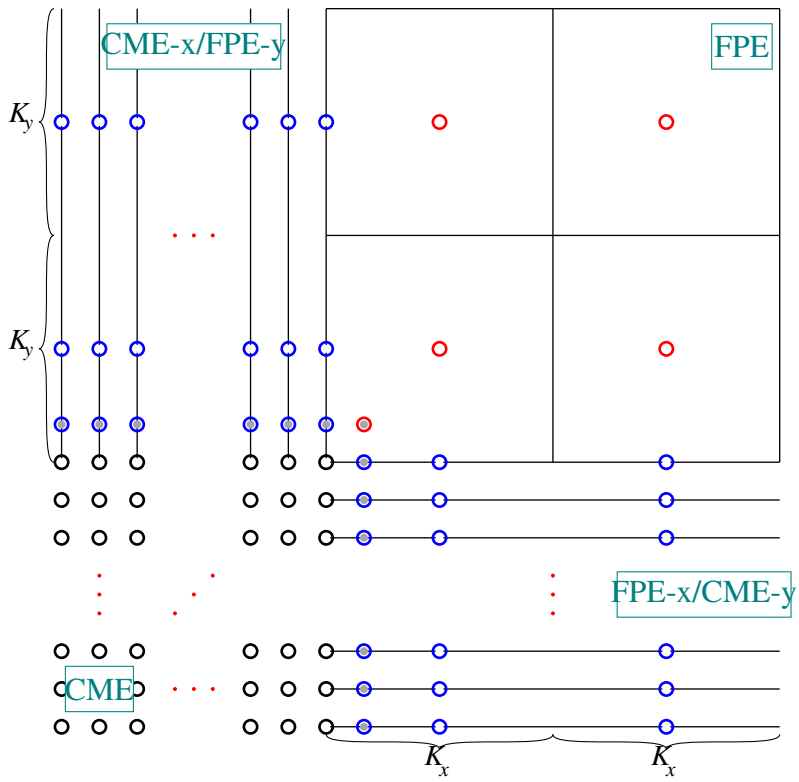

Figure 3.1: Schematic of the CME domain and its interfaces with the hybrid CME-FPE domains and the FPE domain. The open black, blue, and red circles correspond to the CME, hybrid CME-FPE, and FPE states, respectively. The grey-filled circles represent the buffer layer necessary to provide boundary conditions to the CME system.

The time advancement of eq. (3.2) employs a second-order optimal Total Variation Diminishing (TVD) Runge-Kutta scheme [32]:

$$\begin{aligned} p^{(1)} &= p^{(n)} + \Delta t A p^{(n)} \\ p^{(2)} &= p^{(1)} + \Delta t A p^{(1)} \\ p^{(n+1)} &= \frac{1}{2}\left(p^{(n)} + p^{(2)}\right) \end{aligned} \tag{3.12}$$

Here, $p^{(n)}$ is the probability at time t_n, $p^{(n+1)}$ at time $t_{n+1} = t_n + \Delta t$, and A is given in eq. (3.3).

3.4.2 Hybrid CME-FPE

Figure 3.2 shows a schematic of the hybrid FPE-CME domains, labeled with *FPE-x/CME-y* and *CME-x/FPE-y*, and their interfaces with the CME and FPE domains. For an explanation of the color scheme used to represent states in different regions see captions for Fig. 3.1. Additionally, Fig. 3.2 shows states with blue-filled red circles, in a buffer layer required to provide boundary conditions for the hybrid FPE-CME regions. The width of this buffer layer is equal to the largest jump size in the $y-$ and $x-$directions, respectively. The probabilty values are computed by linear interpolation between the adjacent FPE and hybrid FPE-CME states.

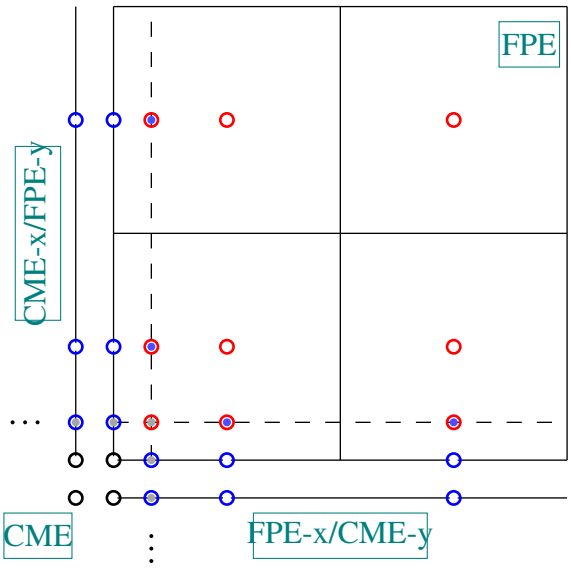

Figure 3.2: Schematic of the hybrid FPE-CME domains and interfaces with the CME and FPE domains.

Figure 3.3 shows a detail of the *FPE-x/CME-y* region, used to illustrate the discretization of eq. (3.10). In this region, state (i, j) is located at $[(i + 1/2)K_x, j]$ in the state space.

The first two terms in the *rhs* of eq. (3.10) are discretized along j lines using a FP approximation,

$$\begin{aligned} &p(x - \boldsymbol{\nu}_r, y - \boldsymbol{\mu}_r)w_r(x - \boldsymbol{\nu}_r, y - \boldsymbol{\mu}_r) - p(x, y - \boldsymbol{\mu}_r)w_r(x, y - \mu_r) \\ &\approx -\xi_{r,x}\frac{\partial}{\partial x}\left(-w_r p(x, y - \mu_r)\right) + \frac{1}{2}\xi_{r,x}^2\frac{\partial}{\partial x}\left(w_r p(x, y - \mu_r)\right). \end{aligned} \tag{3.13}$$

while the last two terms represent vertical "jumps" between states on j lines separated by the appropriate jump sizes μ_r where r is the reaction index. The convection and diffusion terms in the FP approximation (3.13) are discretized as follows

Convection

$$\xi_{r,x}\frac{\partial}{\partial x}\left(w_r p(x, y - \mu_r)\right) \approx \frac{\xi_{r,x}}{K_x}\left(f_{i+1/2}^{(n)} - f_{i-1/2}^{(n)}\right) \tag{3.14}$$

The numerical fluxes $f^{(n)}$ are computed using a 2-nd order essentially non-oscillatory (ENO) reconstruction of probabilties at mid-point locations, and using a Lax-Friedrichs flux-splitting approach to avoid Gibbs phenomena [32]

$$f_{r,i+1/2}^{(n)} = w_{r,i+1/2}\left(p_{i+1/2}^{+} + p_{i+1/2}^{-}\right) + \alpha_{r,i+1/2}\left(p_{i+1/2}^{+} - p_{i+1/2}^{-}\right) \tag{3.15}$$

The probabilities at mid-point locations are constructed as following

$$p_{i+1/2}^{+} = \begin{cases} \frac{3}{2}p_i - \frac{1}{2}p_{i-1} & \text{if } |p_i - p_{i-1}| < |p_{i+1} - p_i|; \\ \frac{1}{2}\left(p_i + p_{i+1}\right) & \text{if } |p_i - p_{i-1}| \geq |p_{i+1} - p_i|. \end{cases} \tag{3.16}$$

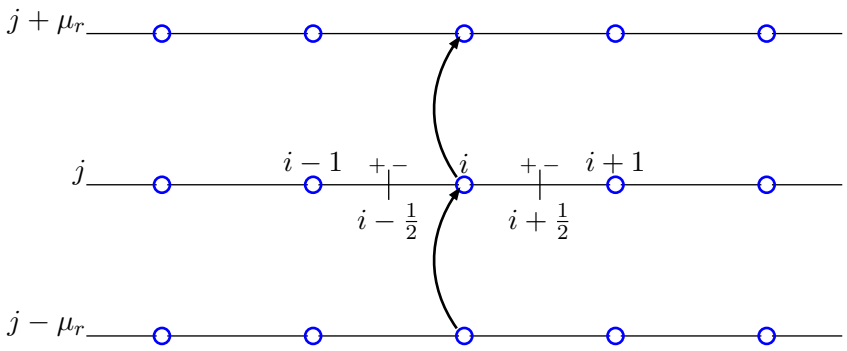

Figure 3.3: Schematic of the hybrid FPE-CME scheme inside the hybrid domain.

The Lax-Friedrichs damping constant α is computed as the maximum propensity w_r for each reaction over a region around the mid-point locations. The effect of the radius of this region, r_α, on the solution is discussed in Section 3.5.

Diffusion

The diffusion term contributions are computed using a 2-nd order finite difference stencil

$$\frac{1}{2}\xi_{r,x}^2\frac{\partial}{\partial x}\left(w_r p(x, y-\mu_r)\right) \approx \frac{2\xi_{r,x}}{K_x}\left(f_{d,i+1/2}^{(n)} - f_{d,i-1/2}^{(n)}\right) \tag{3.17}$$

where

$$f_{d,i+1/2}^{(n)} = \frac{(w_r p)_{i+1} - (w_r p)_i}{K_x}, \qquad f_{d,i-1/2}^{(n)} = \frac{(w_r p)_i - (w_r p)_{i-1}}{K_x} \tag{3.18}$$

The boundary conditions for the combined convection/diffusion flux at the left boundary, $i = 0$, are set to ensure conservation for the transfer of probabilities between the CME domain and the hybrid FPE-CME domain. The domain size is chosen large enough to ensure the probabilities for large counts are small. Therefore, both convection and diffusion fluxes are set to zero on the right boundary corresponding to the sketch shown in Fig. 3.3.

3.4.3 FPE Domain

The schematic in Fig. 3.4 describes the discretization of the convection and diffusion term, in the *rhs* of eq. (3.5). The convection fluxes are discretized in a similar manner as the convection term in the hybrid FPE-CME formulation. The ENO flux-splitting approach is applied sequentially for the x and y components of the convection fluxes.

The diffusion components are based on a conservative approximation using the diffusion fluxes calculated at edge centers, shown with black squares in Fig. 3.4

$$\nabla \cdot \boldsymbol{F}_{r,d} = \xi_{r,x}\frac{\partial}{\partial x}\left(\frac{1}{2}\boldsymbol{\xi}_r \cdot \nabla\left(w_r(\boldsymbol{x})p(\boldsymbol{x};t)\right)\right) + \xi_{r,y}\frac{\partial}{\partial y}\left(\frac{1}{2}\boldsymbol{\xi}_r \cdot \nabla\left(w_r(\boldsymbol{x})p(\boldsymbol{x};t)\right)\right) \tag{3.19}$$

$$\approx \frac{f_{d,i+1/2,j} - f_{d,i-1/2,j}}{K_x} + \frac{f_{d,i,j+1/2} - f_{d,i,j-1/2}}{K_y}. \tag{3.20}$$

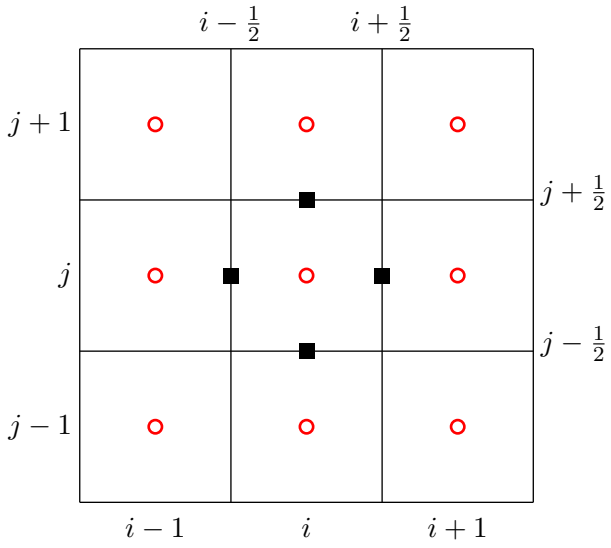

Figure 3.4: Schematic of the computational mesh for the FPE domain. The red circles show the location of the FP states while the black squares show the locations where FP fluxes are computed.

The diffusion fluxes at edge-centers are computed numerically using a 2^{nd} order approximation

$$f_{d,i+1/2,j} = \frac{\xi_{r,x}}{2}\left(\xi_{r,x}\frac{q_{i+1,j}-q_{i,j}}{K_x} + \xi_{r,y}\frac{q_{i+1,j+1}+q_{i,j+1}-q_{i+1,j-1}-q_{i,j-1}}{4K_y}\right) \tag{3.21}$$

where $q = w_r p$.

The boundary conditions for the combined convection/diffusion fluxes at the boundaries with the hybrid CME-FPE domains are set to ensure conservation of probability mass crossing the boundary. Vanishing fluxes are set at far-field boundaries.

3.5 Results and Discussion

Tests were conducted for two kinetic models, both involving two-species systems. The first system involves five reactions modeling the creation of two metabolites, a reaction, and their destruction [7, 9, 36]:

$$\begin{aligned} \emptyset \xrightarrow{k_x} X, \qquad \emptyset \xrightarrow{k_y} Y \\ X + Y \xrightarrow{k_2[x][y]} \emptyset \\ X \xrightarrow{\mu[x]} \emptyset, \qquad Y \xrightarrow{\mu[y]} \emptyset \end{aligned} \tag{3.22}$$

The propensities for the five reactions are given by $w = \{k_x, k_y, k_2 xy, \mu x, \mu y\}^T$, while the jumps $\boldsymbol{\nu}$ (see also Section 3.1) are given by $\{(1,0),(0,1),(-1,-1),(-1,0),(0,-1)\}^T$. We adopted the same values for the model parameters as in [9]: $k_x = k_y = 0.6$, $k_2 = \mu = 0.001$.

The second system models a circadian rhythm using two molecular species, the complex X and the repressor Y, and 4 reactions:

$$\begin{aligned} \emptyset &\xrightarrow{w_1} Y, \quad Y \xrightarrow{w_2} \emptyset \\ Y &\xrightarrow{w_3} X, \quad X \xrightarrow{w_4} Y \end{aligned} \tag{3.23}$$

The expressions for the propensities for the four reactions is provided below for completeness. For more information on the derivation of this model see [36]

$$\begin{aligned} w_1 &= \frac{\beta_R}{\delta_{MR}} \frac{\alpha_R \theta_R + \alpha'_R \gamma_R \tilde{A}(y)}{\theta_R + \gamma_R \tilde{A}(y)} \\ w_2 &= \delta_R y \\ w_3 &= \gamma_C \tilde{A}(y) y \\ w_4 &= \delta_A x \end{aligned} \tag{3.24}$$

where

$$\tilde{A}(y) = \frac{1}{2}\left(\alpha'_A \rho(y) - K_d + \sqrt{\left(\alpha'_A \rho(y) - K_d\right)^2 + 4\alpha_A \rho(y) K_d}\right)$$

and

$$\rho(y) = \frac{\beta_A}{\delta_{MA}(\gamma_C y + \delta_A)}, \;\; K_d = \theta_A / \gamma_A.$$

The coefficients used in the circadian rhythm models are shown in Table 3.1.

α_A	50	β_A	50	γ_A	1	δ_A	1	θ_A	50
α_R	0.01	β_R	5	γ_C	1	δ_R	0.2	θ_R	100
α'_A	500			γ_R	1	δ_{MA}	10		
α'_R	50					δ_{MR}	0.5		

Table 3.1: Coefficients for the circadian rhythm model

3.5.1 Metabolite System: Comparison of FPE to CME

The results presented in this section correspond to simulations using a square computational domain, $0 \leq x \leq 200, 0 \leq y \leq 200$. Figure 3.5 shows a comparison between several FPE simulations, shown with red contours, with the baseline CME simulation shown with black contours. All simulations were started from the same initial condition, a Gaussian blob in the center of the computational domain. These results are used to determine the effect of r_α used in the flux-splitting scheme on the FP approximation. From the results presented in Fig. 3.5, this parameter had little effect on the solution, when $r_\alpha = 2 \ldots 16$ grid points around each computational grid. When the flux-splitting constant α is computed using global maxima for propensities, $r_\alpha = G$, the solution looks significantly diffused, in the lower right frame of Fig. 3.5.

The results shown in Fig. 3.6 are used to investigate the effect of the grid size used in the FP approximation. As the grid size increases, going left to right and top to bottom in the frames, the FPE solution shows increasing discrepancies compared to the CME solution. In other words, while the FPE approximation provides a good solution in areas with many molecules, *i.e.* where discrete

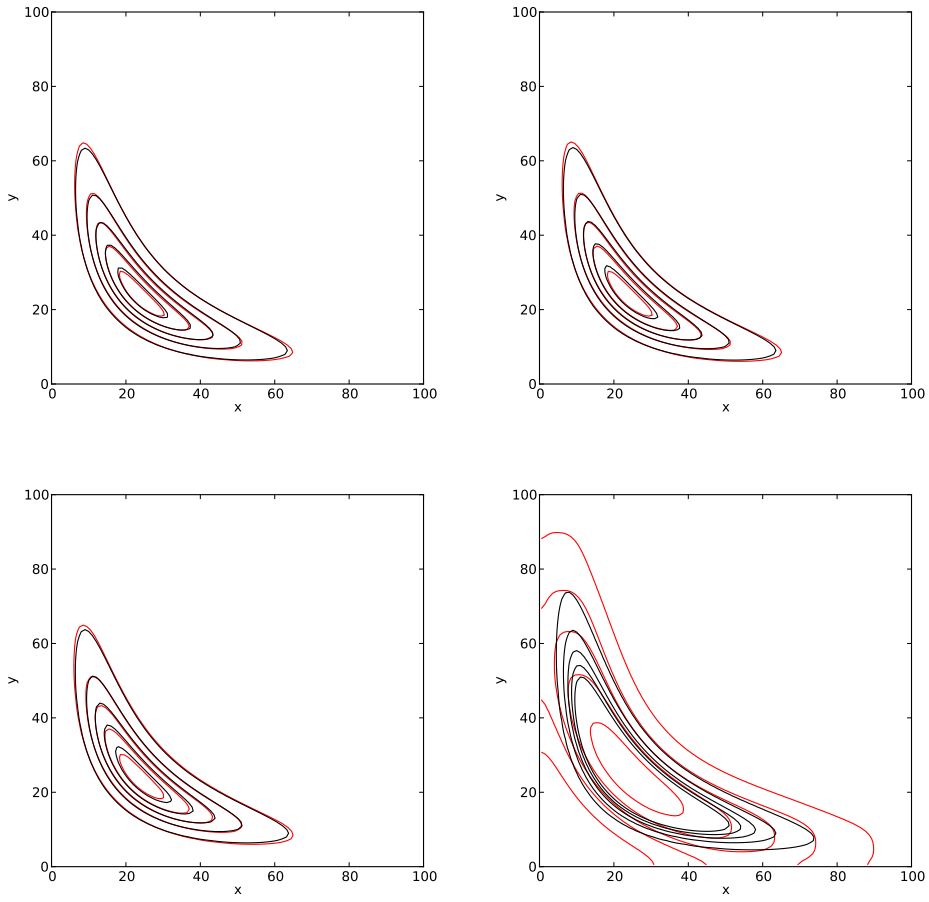

Figure 3.5: Contour plots of CME (black) and FPE (red) solutions at $t = 1500$. Frames are ordered left to right and top to bottom. The FPE solutions correspond to $r_\alpha = 2, 8, 16, G$, respectively. The grid size for all FPE solutions is $\Delta_x = \Delta_y = 1$. The contours levels corrspond to [0.1,0.3,0.5,0.7,0.9] of the maximum FPE value in each frame.

behavior does not matter much, very small grid sizes would be needed to get a good solution in areas with relatively few molecules. For some systems (not shown), the FPE approach does not give reasonable accuracy even for grid sizes less than unity. Therefore, in order to get both good accuracy and efficiency, a combination of CME-FPE approaches needs to be pursued.

Figure 3.7 shows the time evolution of the peak probability as a function of r_α in the left frame and of the grid size in the right frame. Results are consistent with conclusions based on the previous two figures, with the grid size having a strong effect on the time evolution of the solution.

3.5.2 Circadian Rhythm System: FPE Simulations

The solution of the circadian rhythm model is higly sensitive to the numerical approach. The system is cyclic for a number of periods until it reaches a fixed stable point. Figures 3.8 and 3.9 show select snapshots of the solution obtained with the FPE scheme using $r_\alpha = 5$ and $r_\alpha = 10$, respectively. Both

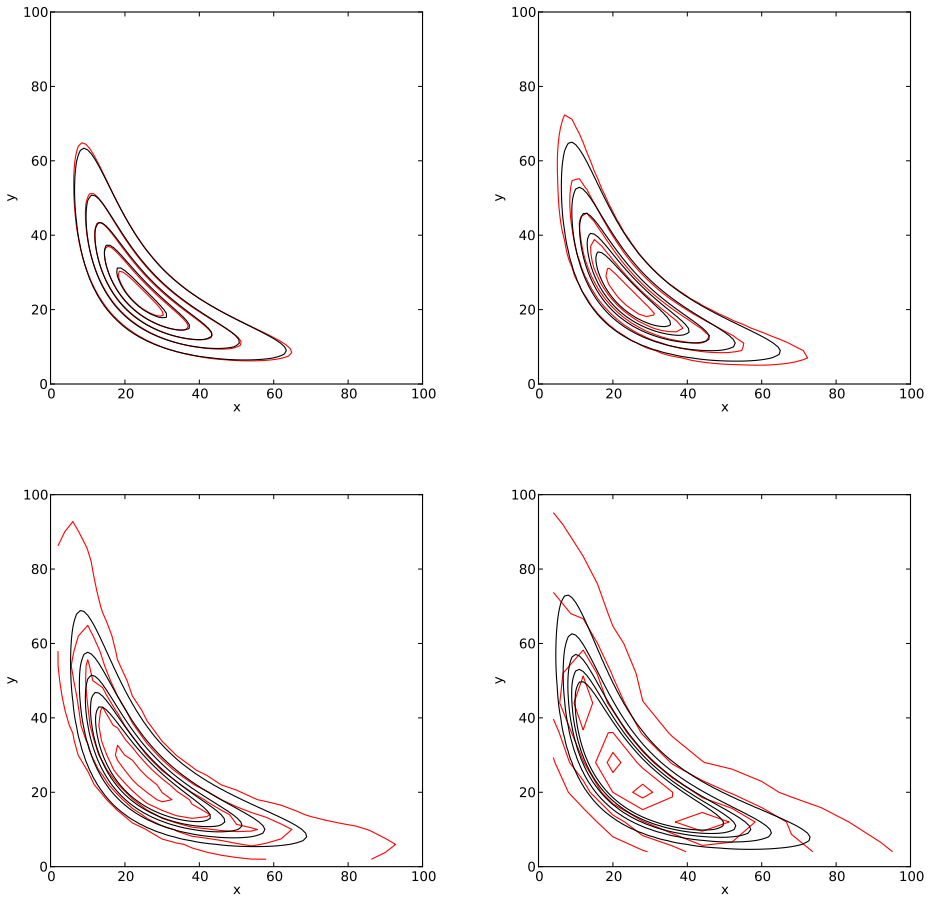

Figure 3.6: Contour plots of CME (black) and FPE (red) solutions at $t = 1500$. Frames are ordered left to right and top to bottom. The FPE solutions correspond to $\Delta_x = \Delta_y = 1, 2, 4, 8$, respectively. The spectral radius for all FPE solutions is $r_\alpha = 2$. The contours levels corrspond to [0.1,0.3,0.5,0.7,0.9] of the maximum FPE value in each frame.

simulations used a computational domain $0 \leq x \leq 1800, 0 \leq y \leq 1800$, a grid size $\Delta_x = \Delta_y = 2$ and the initial condition was a Gaussian blob in the center of the computational domain. These snapshots show that the magnitude of the probability field is significant in several regions of the computational domain. The narrow region near $y = 0$ should be treated with the CME as discrete effects will matter there, while the larger region near $y = 400$ could probably be treated with the continuum FPE. This system should therefore be an appropriate demonstration case for the mixed CME-FPE solution approach once its implementation is fully tested.

Figure 3.10 shows the time evolution of the circadian system modeled with FPE. The results in this figure are used to determine the effect of the spectral radius r_α in the flux-splitting scheme on the cyclical dynamics. In both cases larger windows diminish the amplitude of the oscillations. These observations are consistent with the results for the metabolite system. The grid size has a similar effect, as illustrated in Fig. 3.11. While, eventually the system reaches a steady-state, larger grid sizes and the flux-splitting approach have a dissipative effect and speed up this process.

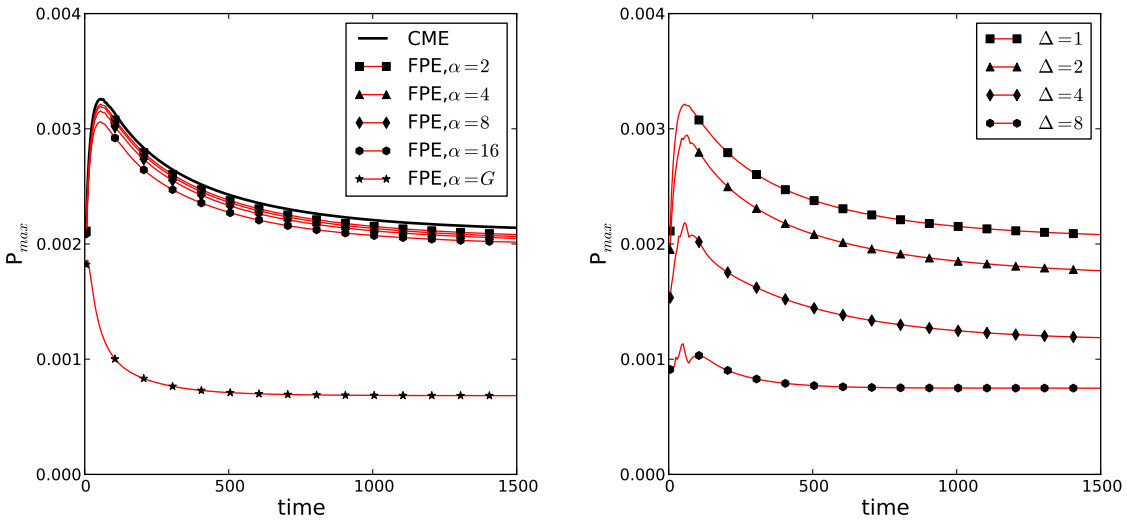

Figure 3.7: Time evolution of maximum probability in the computational domain. The left frame shows FPE results for several spectral radii all for a grid size, $\Delta_x = \Delta_y = 1$. The right frame shows rezults for several grid sizes, all for a spectral radius, $r_\alpha = 2$. The CME solution is shown with a black line in both frames.

3.5.3 Canonical Test of Mixed CME-FPE Formulation

In this section we are showing numerical experiments performed to test the combined mixed CME-FPE algorithms described in Sec. 3.4. These tests are used to investigate the conservation properties of the numerical scheme and to verify its overall order of accuracy.

The computational domain consists of two 2D subdomains, corresponding to the ones marked with **CME** and **FPE-x/CME-y** in Fig. 3.1. The CME subdomain covers $0 \leq x \leq 254$, while the $FPE - x/CME - y$ subdomains extends until $x = 1278$. Both subdomains span the entire $y-$direction, $0 \leq y \leq 511$.

For these canonical tests we considered a kinetic model formed of the first two equations in (3.22)

$$\emptyset \xrightarrow{k_x} X, \quad \emptyset \xrightarrow{k_y} Y \tag{3.25}$$

With the kinetic rates set to constants $k_x = k_y = 6.0$, symmetric in X and Y, and the initial condition for the probability field set to a Gaussian blob centered at $(150, 100)$, the true solution of this system convects the initial probability mass at a 45^o angle to the top-right. Figure 3.12 shows time dynamics of this system for several grid sizes, K_x, in the FPE domain. The first frame, corresponding to the initial condition, is only shown for $K_x = 2$ since all runs share the same starting point.

As the blob crosses the boundary between subdomains, see the second column in Fig. 3.12, the contour lines remain smooth for all grid sizes, verifying that interface conditions between the discrete and continuum formulations are set properly. As expected, at later times the solution corresponding to larger grid sizes is more diffusive. For all grid sizes the solution remains conservative, i.e. $\int_D p\, dv \approx 1$ to within machine precision during the entire simulation time.

We examine the empirical spatial order of accuracy as follows. First, we compute the L_2 norm of

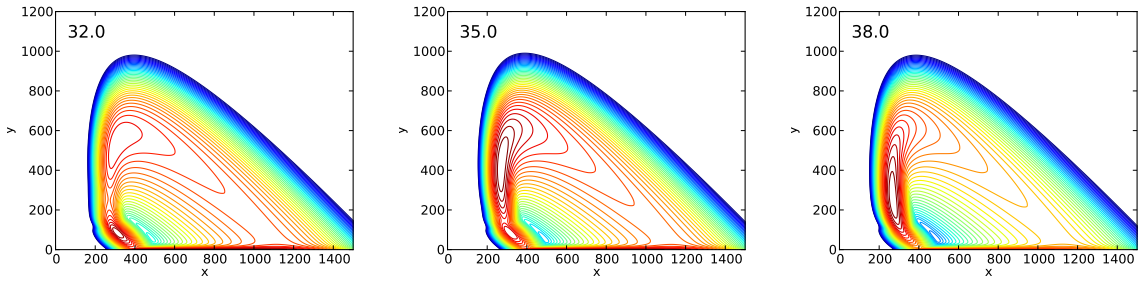

Figure 3.8: Snapshots of the circadian rhythm system modeled using the FP approximation, using $\Delta_x = \Delta_y = 2$, and $r_\alpha = 5$. The time stamp for each snapshot is shown in the upper left corner. The red to blue contour lines are based on the log of probability field, with red contours corresponding to $p \approx 10^{-4}$, and blue contours to $p \approx 10^{-20}$.

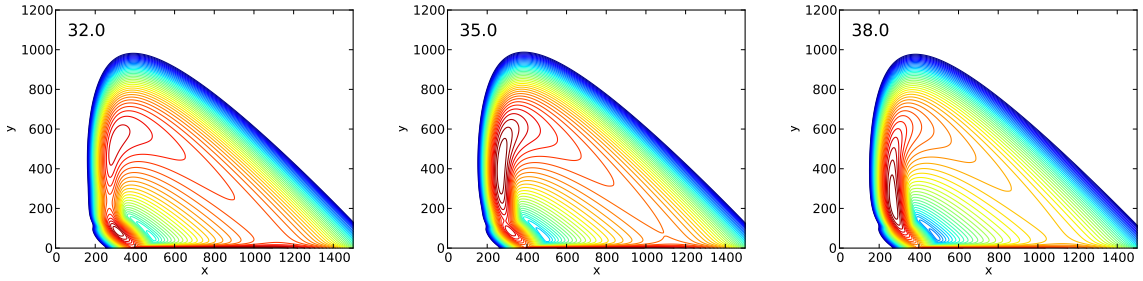

Figure 3.9: Snapshots of the circadian rhythm system modeled using the FP approximation, using $\Delta_x = \Delta_y = 2$, and $r_\alpha = 10$. The time stamp for each snapshot is shown in the upper left corner. The red to blue contour lines are based on the log of probability field, with red contours corresponding to $p \approx 10^{-4}$, and blue contours to $p \approx 10^{-20}$.

the difference between solutions on subsequently refined domains:

$$L_{2,K_x} = \frac{1}{N_{K_x}} \sqrt{\sum_{i,j} (p_{i,j,K_x} - p_{i,j,K_x/2})^2}$$

Here, the sum is taken over all grid points, N_{K_x}, in the simulation corresponding to K_x. The probability values for the simulation using $K_x/2$ are interpolated on the coarser grid corresponding to K_x before computing the L_2 norm. The convergence trend is then computed as

$$2^\gamma = \frac{L_{2,K_x}}{L_{2,K_x/2}} \tag{3.26}$$

The exponent γ is between 1.99 and 2.01 when testing sequences $K_x = \{1, 2, 4\}$ and $K_x = \{2, 4, 8\}$, for several time frames after the solution crossed into the hybrid CME-FPE domain. This verifies that the empirical spatial order of accuracy is 2-nd order, matching the theoretical order of accuracy for the implemented discretization of the hybrid CME-FPE approach.

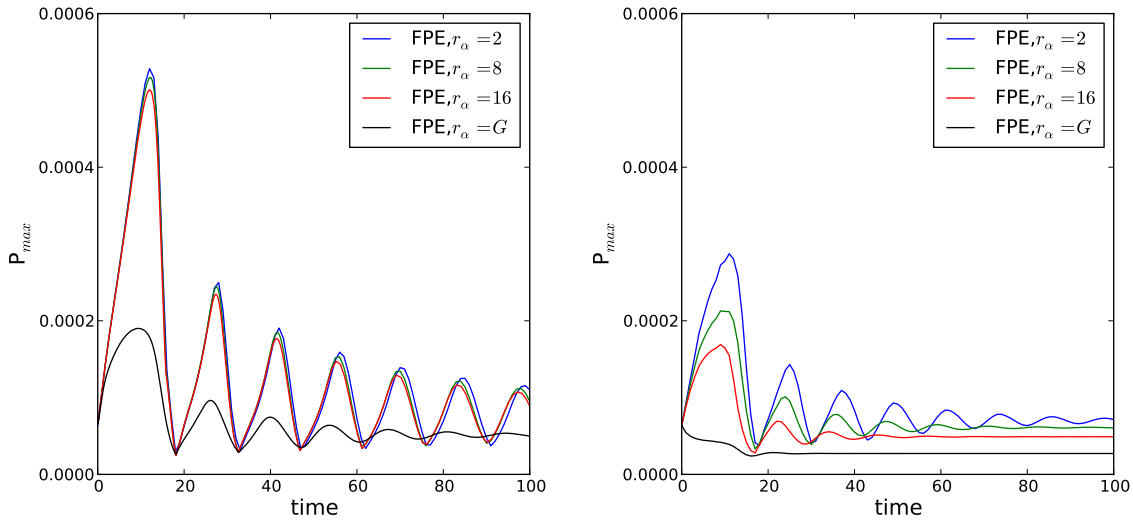

Figure 3.10: Time evolution of maximum probability in the computational domain for several spectral radii, r_α. In left frame the grid size is $\Delta = 2$, while in the right frame, $\Delta = 8$.

The results presented in this section suggest that the proper combination of the CME and FPE approaches presented in Sections 3.1-3.4 will reduce computational times by being able to use larger grid sizes compared to the full CME formulation, while maintaining good accuracy. Tests of the full formulation with 4 2D domains are currently ongoing and the performance of the developed approach will be analyzed on the circadian rhythm system discussed in the previous section.

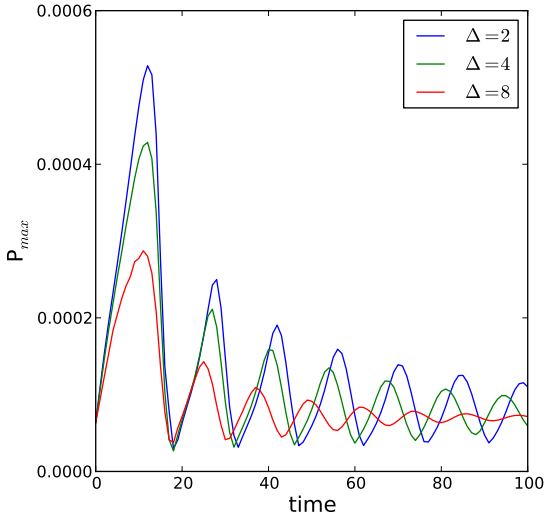

Figure 3.11: Time evolution of maximum probability in the computational domain for several grid sizes. All simulations use $r_\alpha = 8$.

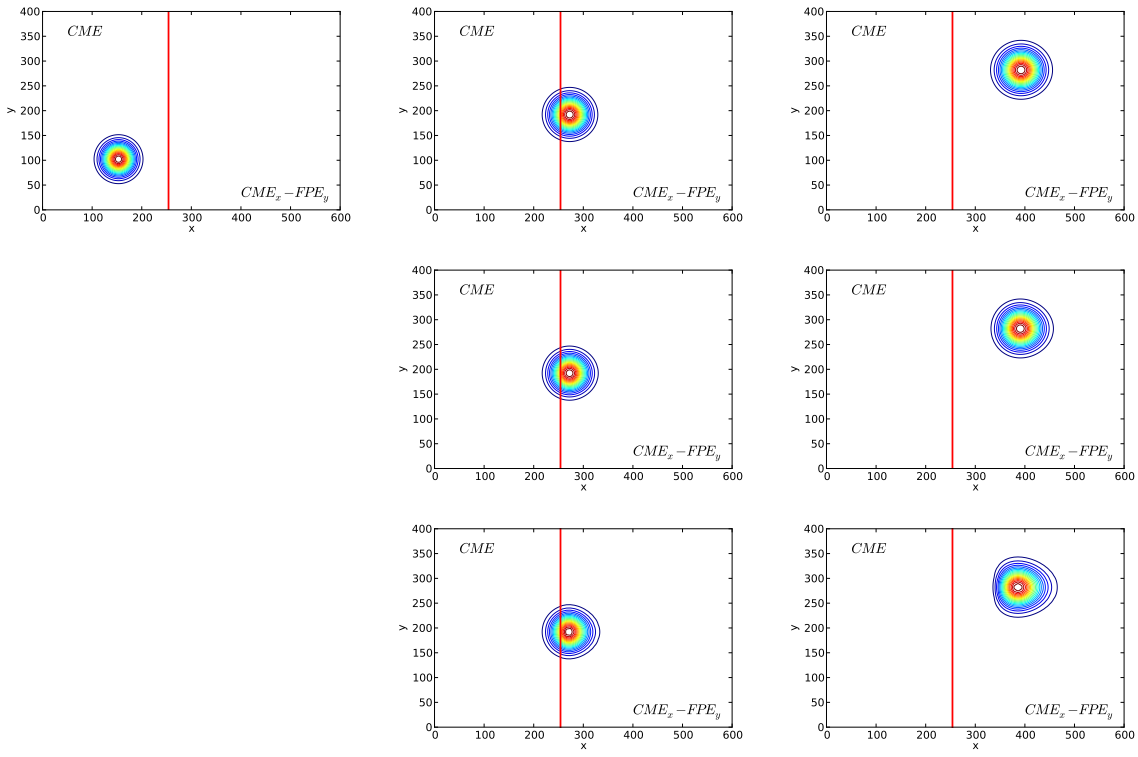

Figure 3.12: Snapshots of probability contours. Solution evolves from the CME domain to the mixed CME-FPE domain. The FPE grid sizes are $K_x = 2, 4$, and 8 for the first, second, and third rows, respectively. The first column corresponds to the initial condition at $t = 0$, second column to $t = 15$, and third column to $t = 30$.

Acknowledgements

This work was supported by the US Department of Energy, Office of Science, Advanced Scientific Computing Research. Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Bibliography

[1] Y. Cao, L. R. Petzold, M. Rathinam, and D. T. Gillespie. The numerical stability of leaping methods for stochastic simulation of chemically reacting systems. *Journal of Chemical Physics*, 121(24):12169 – 12178, Dec 2004.

[2] Yang Cao, Dan Gillespie, and Linda Petzold. Multiscale stochastic simulation algorithm with stochastic partial equilibrium assumption for chemically reacting systems. *Journal of Computational Physics*, 206:395–411, 2005.

[3] Yang Cao, Daniel T. Gillespie, and Linda R. Petzold. The slow-scale stochastic simulation algorithm. *The Journal of Chemical Physics*, 122(1):014116, 2005.

[4] P. Deuflhard, W. Huisinga, T. Jahnke, and M. Wulkow. Adaptive discrete Galerkin methods applied to the chemical master equation. *SIAM J. Sci. Comp.*, 30:2990–3011, 2008.

[5] M.L. Droffner, W.F. Brinton Jr, and E. Evans. Evidence for the prominence of well characterized mesophilic bacteria in thermophilic (50-70°C composting environments. *Biomass and Bioenergy*, 8(3):191–195, 1995.

[6] W. E., D. Liu, and E. Vanden-Eijnden. Nested stochastic simulation algorithms for chemical kinetic systems with multiple time scales. *Journal of Computational Physics*, 221(1):158–180, 2007.

[7] J. Elf, J. Paulsson, O. G. Berg, and M. Ehrenberg. Near-critical phenomena in intracellular metabolite pools. *Biophysical Journal*, 84:154–170, 2003.

[8] S. Engblom. A discrete spectral method for the chemical master equation. Technical Report 36, Uppsala University, 2006.

[9] L. Ferm, P. Lötstedt, and P. Sjöberg. Adaptive, conservative solution of the Fokker-Planck equation in molecular biology. Technical Report 2004-054, Department of Information Technology, Uppsala University, Uppsala, Sweden, 2004.

[10] M.A. Gibson and J. Bruck. Efficient Exact Stochastic Simulation of Chemical Systems with Many Species and Many Channels. *J. Phys. Chem. A*, 104:1876–1889, 2000.

[11] D. T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *Journal of Chemical Physics*, 115(4):1716 – 33, Jul 2001.

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

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

[14] Drew Gorman-Lewis, Patricia E. Elias, and Jeremy B. Fein. Adsorption of aqueous uranyl complexes onto bacillus subtilis cells. *Environmental Science and Technology*, 39(13):4906–4912, 2005.

[15] R. Gunawan, Y. Cao, L. Petzold, and F. J. Doyle. Sensitivity analysis of discrete stochastic systems. *Biophysical Journal*, 88(4):2530 – 2540, Apr 2005.

[16] Ryan N. Gutenkunst, Joshua J. Waterfall, Fergal P. Casey, Kevin S. Brown, Christopher R. Myers, and James P. Sethna. Universally sloppy parameter sensitivities in systems biology models. *PLOS Computational Biology*, 3(10):1871–1878, 2007.

[17] E. L. Haseltine and J. B. Rawlings. Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *Journal of Chemical Physics*, 117(15):6959 – 6969, Oct 2002.

[18] M. Hegland, C. Burden, L. Santoso, S. Macnamara, and H. Booth. A solver for the stochastic master equation applied to gene regulatory networks. *J. of Comp. and Appl. Math.*, 205:708–724, 2007.

[19] D. Kim, B.J. Debusschere, and H.N. Najm. Spectral methods for parametric sensitivity in stochastic dynamical systems. *Biophys. Journal*, 92(2):379 – 393, Jan 2007.

[20] S. MacNamara, A.M. Bersani, K. Burrage, and R.B. 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*, 129:95–105, 2008.

[21] 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.

[22] B. Munsky and M. Khammash. The finite state projection algorithm for the solution of the chemical master equation. *Journal of Chemical Physics*, 124(4):44104–1–13, 2006.

[23] 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.

[24] K.S. Noah, S.L. Fox, D.F. Bruhn, D.N. Thompson, and G.A. Bala. Development of Continuous Surfactin Production from Potato-Process Effluent by *Bacillus subtilis* in an Airlift Reactor. *Applied Biochemistry and Biotechnology*, 98-100:803–813, 2002.

[25] S. Plyasunov and A.P. Arkin. Efficient stochastic sensitivity analysis of discrete event systems. *Journal of Computational Physics*, 2006. in press.

[26] M. Rathinam, L. R. Petzold, Y. Cao, and D. T. Gillespie. Stiffness in stochastic chemically reacting systems: the implicit tau-leaping method. *Journal of Chemical Physics*, 119(24):12784 – 94, DEC 2003.

[27] K. Sargsyan, B. Debusschere, H. Najm, and O. Le Maître. Spectral representation and reduced order modeling of the dynamics of stochastic reaction networks via adaptive data partitioning. *SIAM Journal on Scientific Computing*, 31(6):4395–4421, 2010.

[28] 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):2283–2297, 2009.

[29] K. Sargsyan, C. Safta, B. Debusschere, and H. Najm. Multiparameter spectral representation of noise-induced competence in Bacillus subtilis. *IEEE/ACM Trans. Comp. Biol. and Bioinf.*, 9, 2012.

[30] K.D. Schaller, S.L. Fox, D.F. Bruhn, K.S. Noah, and G.A. Bala. Characterization of Surfactin from *Bacillus subtilis* for Application as an Agent for Enhanced Oil Recovery. *Applied Biochemistry and Biotechnology*, 113-116:827–836, 2004.

[31] Henning Schmidt, Mads F. Madsen, Sune Dano, and Gunnar Cedersund. Complexity reduction of biochemical rate expressions. *Bioinformatics*, 24(6):848–854, 2008.

[32] Chi-Wang Shu. Essentially Non-Oscillatory and Weighted Essentially Non-Oscillatory Schemes for Hyperbolic Conservation Laws. Technical report, ICASE Report No. 97-65; NASA/CR-97-206253, November 1997.

[33] R.B. Sidje, K. Burrage, and S. MacNamara. Inexact uniformization method for computing transient distributions of markov chains. *SIAM Journal on Scientific Computing*, 29(6):2562–80, 2007.

[34] P. Sjöberg. PDE and Monte Carlo approaches to solving the master equation applied to gene regulation. Technical Report 2007-028, Uppsala University, http://www.it.uu.se/research/publications/reports/2007-028/2007-028-nc.pdf, 2007.

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

[36] J. M. G. Vilar, H. Y. Kueh, N. Barkai, and S. Leibler. Mechanisms of noise-resistance in genetic oscillators. *Proc. Nat. Acad. Sci.*, 99:5988–5992, 2002.

Appendix A

Multiparameter spectral representation of noise-induced competence in Bacillus subtilis

Multiparameter Spectral Representation of Noise-Induced Competence in Bacillus Subtilis

Khachik Sargsyan, Cosmin Safta, Bert Debusschere, and Habib Najm

Abstract—In this work, the problem of representing a stochastic forward model output with respect to a large number of input parameters is considered. The methodology is applied to a stochastic reaction network of competence dynamics in *Bacillus subtilis* bacterium. In particular, the dependence of the competence state on rate constants of underlying reactions is investigated. We base our methodology on Polynomial Chaos (PC) spectral expansions that allow effective propagation of input parameter uncertainties to outputs of interest. Given a number of forward model training runs at sampled input parameter values, the PC modes are estimated using a Bayesian framework. As an outcome, these PC modes are described with posterior probability distributions. The resulting expansion can be regarded as an uncertain response function and can further be used as a computationally inexpensive surrogate instead of the original reaction model for subsequent analyses such as calibration or optimization studies. Furthermore, the methodology is enhanced with a classification-based mixture PC formulation that overcomes the difficulties associated with representing potentially nonsmooth input-output relationships. Finally, the global sensitivity analysis based on the multiparameter spectral representation of an observable of interest provides biological insight and reveals the most important reactions and their couplings for the competence dynamics.

Index Terms—Approximation, spectral methods, probability and statistics.

1 Introduction

Biochemical models are most commonly described by rate equations, i.e., by a system of ordinary differential equations (ODEs) governing the time evolution of species concentrations. This is merely a *macroscopic* approximation that is accurate only if the relevant volume is sufficiently large. For small volumes or small species numbers the intrinsic stochastic noise due to random molecular collisions becomes more significant, and a microscopic description becomes necessary. Stochastic reaction networks (SRNs) provide such a framework, representing the time evolution of the species numbers as a discrete state, continuous-time Markov jump process [1], [2]. Despite recent advances in solving the governing chemical master equations (CMEs) for the probability distributions of the species numbers (see, e.g., [3], [4], [5]), simulation-based methods still serve as the main analytical tools for analyzing the dynamics of the system and their sensitivity with respect to the input reaction rates. Specifically, the stochastic simulation algorithm [6], [7] (SSA) provides a mechanism for the time evolution of species numbers, effectively sampling the CME solution and allowing its statistical analysis. One can obtain statistical estimates of output quantities of interest by sampling several SSA realizations.

Input parameters of stochastic reaction networks are the rate constants and other auxiliary parameters involved in the specification of reaction propensities. Usually, these parameters are estimated empirically and can therefore have large uncertainties. As a result, it is essential to be able to perform a global sensitivity analysis, where large ranges of parameter perturbations and their effect on the output quantities of interest are investigated. Most of the literature up to now has focused on local, derivative-based sensitivity in stochastic reaction networks [8], [9], [10]. In this work, we develop a rigorous approach to build a response surface approximation of the output observables with respect to the input parameters. This response surface can serve as a surrogate model and be queried instead of the full model in studies that require prohibitively many simulations. It enables analyses such as global sensitivity analysis, uncertainty quantification (UQ) and propagation, optimization, as well as efficient calibration of input parameters against available experimental data associated with the model outputs. For example, when experimental data are available, one can use Bayesian techniques and Markov chain Monte Carlo (MCMC) sampling to infer posterior probability distributions for rate constants. However, MCMC generally requires prohibitively many model simulations in order to obtain adequate posterior samples. The response surface therefore can be invoked instead as a computationally affordable surrogate for the full model [11], [12].

The benchmark system in this work is the reaction network of competence transition in *Bacillus subtilis* [13]. Competence is a state of a cell that allows it to take DNA from the environment, as opposed to a vegetative state. It is generally believed that the transition from a vegetative state to a competent one is driven by the intrinsic noise in the system [13], [14], [15], [16]. However, these studies are typically performed at fixed-parameter settings and do not generally cover uncertainties associated with the lack-of-knowledge of reaction rates. At least one group has looked at large parameter perturbations [13], [14] analyzing, however, only a couple of parameters at a time. Our goal is to investigate the

• *The authors are with Sandia National Laboratories, 7011 East Ave., MS 9051, Livermore, CA 94550.*
E-mail: {ksargsy, csafta, bjdebus, hnnajm}@sandia.gov.

Manuscript received 30 Jan. 2012; revised 11 June 2012; accepted 17 July 2012; published online 1 Aug. 2012.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2012-01-0023. Digital Object Identifier no. 10.1109/TCBB.2012.107.

reaction network at hand by building a multiparameter surrogate or response surface that approximates the input-output relationship over the range of variation of the parameters and is queried for global sensitivity analysis.

There are several challenges that generic uncertainty quantification or response surface methods face when dealing with SRNs in general and the *B.subtilis* competence network in particular. First of all, many methods rely on smoothness of the input-output relationship or forward function. Therefore, the presence of leftover noise, even after averaging over many realizations, renders these methods infeasible in their conventional form. Similarly, the nonlinear character of the forward function strongly affects the convergence properties of smooth-basis representations. Furthermore, reaction networks with many reactions have a large number of rate constants or input parameters. Again, since most of these parameters are uncertain, one would like to perform uncertainty and sensitivity analysis considering all these parameters. Inevitably, the curse of dimensionality becomes an important issue—in order to explore the input parameter space sufficiently well, one needs to sample SSA realizations in prohibitively many parameter regimes.

In this paper, we develop and implement a Bayesian strategy to obtain a multidimensional response surface representation of noisy forward functions given a limited amount of *training* runs corresponding to a sampled set of inputs. First of all, we rely on Polynomial Chaos (PC) spectral expansions to represent the forward function as an expansion in a polynomial basis. At this stage, we effectively have a polynomial response surface fit, while the PC framework allows for subsequent analysis with nontrivial input parameter distributions. We extend previous work [17] to include larger parameter variabilities in order to perform global sensitivity assessment. Also, while employed in a lower dimensional setting in [17], orthogonal projection on the polynomial basis in high dimensions suggests the use of sparse quadrature methods to alleviate the curse of dimensionality. However, with noisy forward function evaluations, these methods are expected to fail due to a lack of sufficient smoothness in the projected function. Therefore, we adopt a Bayesian approach, where the PC coefficients are inferred using the available, albeit noisy, function evaluations. The Bayesian framework [18] is particularly well suited for fitting problems with different sources of noise and uncertainties. In this case, besides parametric variability, there are also the intrinsic noise of the system, as well as uncertainties associated with the lack-of-sampling, i.e., with *sparsity* of the training set. The effect of these uncertainties is reflected in the posterior distribution of the PC coefficients inferred from the training run data, effectively leading to an *uncertain response surface* for the output observable dependence on the input parameters. We use conjugate priors to arrive at analytically tractable posterior distributions.

The paper is organized as follows: Section 2 describes the benchmark problem of competence dynamics in *B.subtilis* with the noise-induced competence demonstration. Sections 3 and 4 are devoted to the general formulation of the problem and the techniques we tackle the problem with. Next, Section 5 describes some of the challenges associated with the methods and the proposed improvements. Finally, Section 6 focuses on the results, i.e., a 2D response surface construction as a proof-of-concept and a higher dimensional case involving the full set of the rate constants, enhanced by a sensitivity-based parameter down selection. We then draw some conclusions in Section 7.

2 Stochastic Reaction Network of Competence in *Bacillus Subtilis*: ODE Limit and Noise-Induced Competence

We explore the stochastic reaction network for competence in *Bacillus subtilis*, which is a gram-positive soil bacterium. Competence in this bacterium is a state that allows uptake of external DNA, letting *B. subtilis* incorporate new genetic material and thus increase its fitness [13], [15]. The reaction network consists of 11 species and 16 reactions, summarized in Tables 1 and 2, see [13]. The reaction network is depicted in Fig. 1. We will sample the process using Gillespie's Stochastic Simulation Algorithm [6], [7], which provides a practical mechanism of simulating the time evolution of species molecule counts. The software package StochKit2 has been used for SSA simulations [19] . The transcription factor comK can be considered as an indicator of competence. Competence is characterized by large quantities of comK, as demonstrated in Fig. 2A. Furthermore, Fig. 2B shows sampled phase space trajectories of two species, comK and comS. Clearly, as can be seen from the log-plots, at the nominal parameter setting, there are two basins of attraction in the phase space. The system spends most of the time around the vegetative state with low number of comK molecules with sporadic excursions to high values, corresponding to the competent state.

In the limit of large system volume Ω, the intrinsic noise can be neglected and the reaction network dynamics can be approximated by an Ordinary Differential Equation (ODE) system for the concentrations of the species. For comK and comS concentrations, denoted by K and S, respectively, the rescaled and dimensionless versions of the equations are

$$\begin{aligned}\frac{1}{t_K}\frac{dK}{dt} &= a_k + \frac{b_k K^n}{\kappa_0^n + K^n} - \frac{K}{1+K+S} - \Delta_k K,\\ \frac{1}{t_S}\frac{dS}{dt} &= a_s + \frac{b_s k_1^p}{\kappa_1^p + K^p} - \frac{S}{1+K+S} - \Delta_s S,\end{aligned} \tag{1}$$

with the time rescaling factors

$$t_K = \frac{k_{-11}+k_{12}}{N_{\mathrm{MecA}} k_{11} k_{12}}, \qquad t_S = \frac{k_{-13}+k_{14}}{N_{\mathrm{MecA}} k_{13} k_{14}}, \tag{2}$$

while the rescaled concentrations are

$$K \to \frac{\Omega k_{11}}{k_{-11}+k_{12}} K, \qquad S \to \frac{\Omega k_{13}}{k_{-13}+k_{14}} S, \tag{3}$$

and the dimensionless parameters are [13]

$$\begin{aligned} a_k &= \frac{k_1 k_3}{k_7 k_{12}} \frac{N_{\mathrm{P}_{\mathrm{comK}}^{\mathrm{const}}}}{N_{\mathrm{MecA}}}, & b_k &= \frac{k_2 k_3}{k_7 k_{14}} \frac{N_{P_{\mathrm{comK}}}}{N_{\mathrm{MecA}}},\\ a_s &= \frac{k_4 k_6}{k_9 k_{14}} \frac{N_{\mathrm{P}_{\mathrm{comS}}^{\mathrm{const}}}}{N_{\mathrm{MecA}}}, & b_s &= \frac{k_5 k_6}{k_9 k_{14}} \frac{N_{\mathrm{P}_{\mathrm{comS}}}}{N_{\mathrm{MecA}}},\\ \kappa_0 &= \frac{k_k k_{11}}{k_{-11}+k_{12}}, & \kappa_1 &= \frac{k_s k_{13}}{k_{-13}+k_{14}},\\ \Delta_k &= \frac{k_8(k_{-11}+k_{12})}{k_{11}k_{12}}, & \Delta_s &= \frac{k_{10}(k_{-13}+k_{14})}{k_{13}k_{14}}. \end{aligned} \tag{4}$$

TABLE 1
The Set of Reactions in the Model

Reaction	Parameters	Nominal values [13]
$\mathrm{P}_{\mathrm{comK}}^{\mathrm{const}} \xrightarrow{k_1} \mathrm{P}_{\mathrm{comK}}^{\mathrm{const}} + \mathrm{mRNA}_{\mathrm{comK}}$	k_1	$0.00021875\ \mathrm{s}^{-1}$
$\mathrm{P}_{\mathrm{comK}} \xrightarrow{f(K,k_2,k_k,n)} \mathrm{P}_{\mathrm{comK}} + \mathrm{mRNA}_{\mathrm{comK}}$	k_2	$0.1875\ \mathrm{s}^{-1}$
	k_k	$5000\ \mathrm{nM}$
	n	2
$\mathrm{mRNA}_{\mathrm{comK}} \xrightarrow{k_3} \mathrm{mRNA}_{\mathrm{comK}} + \mathrm{comK}$	k_3	$0.2\ \mathrm{s}^{-1}$
$\mathrm{P}_{\mathrm{comS}}^{\mathrm{const}} \xrightarrow{k_4} \mathrm{P}_{\mathrm{comS}}^{\mathrm{const}} + \mathrm{mRNA}_{\mathrm{comS}}$	k_4	$0.0\ \mathrm{s}^{-1}$ *
$\mathrm{P}_{\mathrm{comS}} \xrightarrow{g(K,k_5,k_s,p)} \mathrm{P}_{\mathrm{comS}} + \mathrm{mRNA}_{\mathrm{comS}}$	k_5	$0.0015\ \mathrm{s}^{-1}$
	k_s	$833\ \mathrm{nM}$
	p	5
$\mathrm{mRNA}_{\mathrm{comS}} \xrightarrow{k_6} \mathrm{mRNA}_{\mathrm{comS}} + \mathrm{comS}$	k_6	$0.2\ \mathrm{s}^{-1}$
$\mathrm{mRNA}_{\mathrm{comK}} \xrightarrow{k_7} \emptyset$	k_7	$0.005\ \mathrm{s}^{-1}$
$\mathrm{comK} \xrightarrow{k_8} \emptyset$	k_8	$0.0001\ \mathrm{s}^{-1}$
$\mathrm{mRNA}_{\mathrm{comS}} \xrightarrow{k_9} \emptyset$	k_9	$0.005\ \mathrm{s}^{-1}$
$\mathrm{comS} \xrightarrow{k_{10}} \emptyset$	k_{10}	$0.0001\ \mathrm{s}^{-1}$
$\mathrm{MecA} + \mathrm{comK} \xrightarrow{k_{11}/\Omega} \{\mathrm{MecA}\vert\mathrm{comK}\}$	k_{11}	$2.02 \cdot 10^{-6}\ \mathrm{s}^{-1}$
$\mathrm{MecA} + \mathrm{comK} \xleftarrow{k_{-11}} \{\mathrm{MecA}\vert\mathrm{comK}\}$	k_{-11}	$0.0005\ \mathrm{s}^{-1}$
$\{\mathrm{MecA}\vert\mathrm{comK}\} \xrightarrow{k_{12}} \mathrm{MecA}$	k_{12}	$0.05\ \mathrm{s}^{-1}$
$\mathrm{MecA} + \mathrm{comS} \xrightarrow{k_{13}/\Omega} \{\mathrm{MecA}\vert\mathrm{comS}\}$	k_{13}	$4.5 \cdot 10^{-6}\ \mathrm{s}^{-1}$
$\mathrm{MecA} + \mathrm{comS} \xleftarrow{k_{-13}} \{\mathrm{MecA}\vert\mathrm{comS}\}$	k_{-13}	$0.00005\ \mathrm{s}^{-1}$
$\{\mathrm{MecA}\vert\mathrm{comS}\} \xrightarrow{k_{14}} \mathrm{MecA}$	k_{14}	$0.00004\ \mathrm{s}^{-1}$

*Ω is a factor, proportional to the volume and is set to $\Omega = 1\ \mathrm{nM}^{-1}$. The Hill functions $f(K,k_2,k_k,n) = \frac{k_2 K^n}{k_k^n + K^n}$ and $g(K,k_3,k_s,p) = \frac{k_3 k_s^p}{k_s^p + K^p}$, where K is the comK molecule-count, approximate the transcription dynamics of comK and comS, respectively, bulking several component reactions into a single one. *The nominal value in our studies is adjusted to $0.001 s^{-1}$ in order to enable two-sided parameter domain exploration.*

We denoted by N_X the total number of molecules X (in case of MecA, this includes the complexes formed with comK and comS). It is clear from the reactions that the total molecule-counts of $\mathrm{P}_{\mathrm{comK}}^{\mathrm{const}}$, $\mathrm{P}_{\mathrm{comS}}^{\mathrm{const}}$, $\mathrm{P}_{\mathrm{comK}}$, $\mathrm{P}_{\mathrm{comS}}$, and MecA are conserved quantities as no reaction alters their quantities. Also note that in the nominal setting $t_K = t_S$, i.e., comK and comS vary at the same scale.

Next, in order to separate out the large-scale dynamics of the system from smaller scale fluctuations that are due to the intrinsic noise, we will change parameters in the discrete stochastic system in a way that keeps the underlying ODE dynamics the same. For example, one can change the parameters k_1 (constitutive transcription of comK), k_2 (regulated transcription of comK), and k_3 (translation of comK) such that the products $k_1 k_3$ and $k_2 k_3$ remain constant, ensuring that the corresponding ODE equations remain the same; see (4). If one increases the translation rate k_3 by a factor and reduces the transcription rates k_1 and k_2 by the same factor, the ODE dynamics remains the same, while the noise level in the system increases. In other words, transcription and translation balance the large scale ODE dynamics, while making a difference to the small scale fluctuations. On the logarithmic scale this corresponds to the exploration of the 3D input parameter space $(\log k_1, \log k_2, \log k_3)$ along the line

$$\log k_3 = C_1 - \log k_1 = C_2 - \log k_2, \tag{5}$$

TABLE 2
Description of the Species Involved in the Reaction Network, as Well as the Initial Molecule-Counts Used for SSA Simulations

Species	Description	Initial molecule-count [13]
comK	master regulator	25
comS	small peptide	200
$\mathrm{mRNA_{comK}}$	messenger RNA for comK	1
$\mathrm{mRNA_{comS}}$	messenger RNA for comS	1
$\mathrm{P_{comK}}$	regular promoter of comK	1
$\mathrm{P_{comS}}$	regular promoter of comS	1
$\mathrm{P_{comK}^{const}}$	constitutive promoter of comK	1
$\mathrm{P_{comS}^{const}}$	constitutive promoter of comS	1
MecA	protease	500
$\{\mathrm{MecA}\vert\mathrm{comK}\}$	protease complex	0
$\{\mathrm{MecA}\vert\mathrm{comS}\}$	protease complex	0

Note that the total concentration of promoters, messengers, as well as protease MecA, stay constant according to the reaction network described in Table 1.

for some constants C_1 and C_2. Specifically, these constants were chosen to ensure the line goes through the nominal values, i.e., $C_i = \log \tilde{k}_i + \log \tilde{k}_3$, for $i = 1, 2$. Here and thereafter, the nominal value of a parameter will be denoted by a tilde notation.

Fig. 3 shows the trajectories for three parameter settings along the line described by (5). In particular, we illustrate the nominal setting and two settings with higher or lower intrinsic noise, corresponding to a change in k_3 by a factor of 1.2. These settings share the same ODE dynamics, also illustrated in Fig. 3, while the increased noise level generates more transitions to competence due to sporadic large jumps of the comK molecule-counts into the competence regime, generally confirming conclusions found in the literature [13].

The perturbation analysis in this section corresponds to tailored parameter modifications, since parameters were perturbed along the line from (5) in a 3D space. This was done primarily to emphasize the effect of intrinsic noise in the system. However, given reasonable domains for perturbation of each parameter, one needs to carry out a full predictability analysis in the parameter space that accounts for all possible parameter combinations, i.e., in a hypercube that is the product space of single parameter domains. In the next section, we develop a general strategy for multiparameter response surface construction that

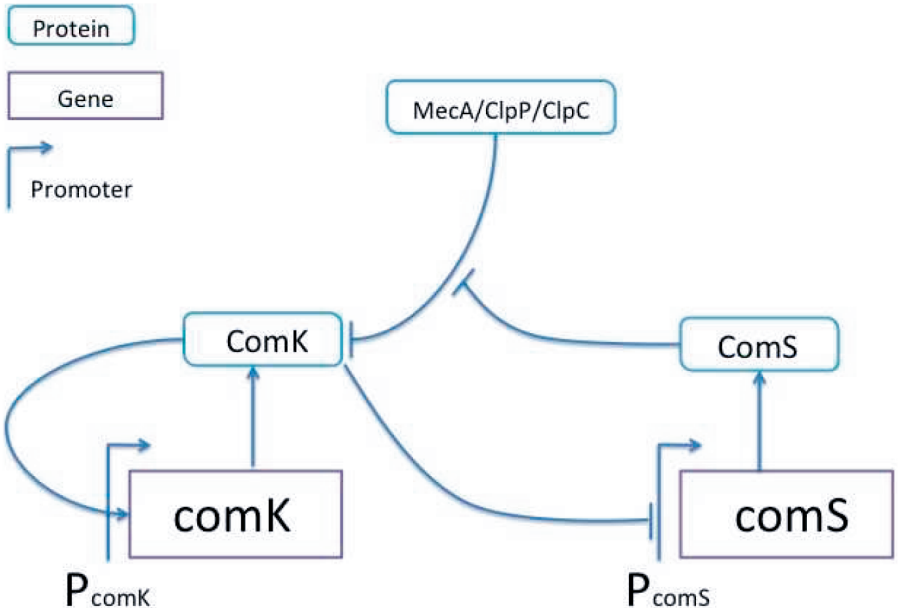

Fig. 1. The reaction network of the competence dynamics in $B.subtilis$. Species ComS and ComK compete for degradation by the MecA complex. Also note the positive transcriptional feedback loop of comK and the negative feedback loop in which ComK indirectly inhibits expression of ComS.

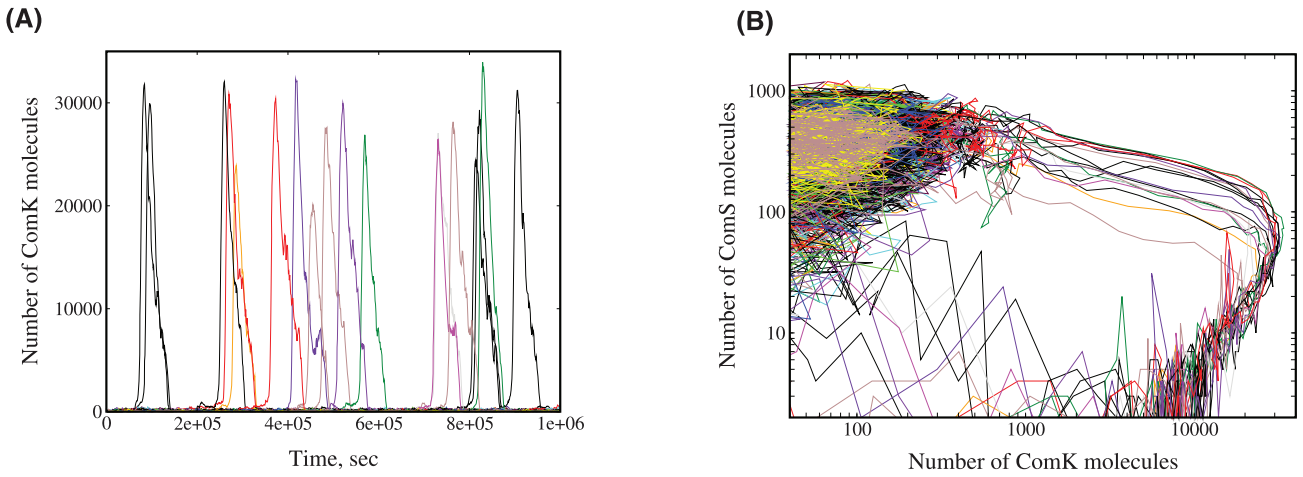

Fig. 2. A hundred sample realizations at nominal parameter values. (A) The number of comK molecules versus time, (B) The phase space of comK and comS molecule-counts.

allows efficient exploration of the full, high-dimensional parametric space.

3 Polynomial Chaos Response Surface Construction for Competence Probability

Let us take the probability of competence P_c in the steady state (i.e., excluding the initial transient period) as our output quantity of interest. Specifically, we choose a threshold value of 5,000 and assume that competence is characterized by $K = N_{comK} > 5{,}000$ where K is the number of comK molecules. Writing the time dependence explicitly, we can estimate the true value of P_c by

$$P_c \approx Z = \frac{1}{t_f - t_s}\int_{t_s}^{t_f} \mathbb{I}_{K(t)>5{,}000}\,dt, \tag{6}$$

where $\mathbb{I}_{\mathcal{S}}$ is equal to one if the statement $\mathcal{S}$ is true, and vanishes otherwise. For each replica simulation, the quantity in (6) is computed as the fraction of time the comK molecule-count $K(t)$ is above the threshold 5,000, between the starting time t_s to lose the initial condition dependence, and the final time t_f. Unless otherwise noted, these time values are set to $t_s = 2 \cdot 10^5$ sec and $t_f = 10^6$ sec. Each realization will produce an estimate Z of the true probability of competence P_c, the discrepancy being due to both intrinsic stochasticity and the finite time window. In parameter regimes with frequent competence events, the finite time window (t_s, t_f) is sufficiently large to estimate the competence fraction reliably. In parameter regions with rare competence events, the chosen time window may not be large enough to reliably estimate the competence probability in each individual realization. However, as we describe below, replicating the simulations and averaging the competence time fraction over an ensemble of many time windows effectively enlarges the overall time horizon and produces more reliable competence probability estimates.

PC spectral expansions [20], [21] will be employed to represent the dependence of P_c on the input parameter vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$ consisting of reaction rate constants. In the present context, $\boldsymbol{\lambda}$ is treated as a random vector with independent components that have arbitrary distributions. We approximate both input and output random variables in terms of an expansion with respect to an orthogonal set of polynomials of specific standard random variables. In this work, we will employ Legendre-Uniform (LU) PC expansions. The LU PC expansions can simply be interpreted as a response surface or a polynomial fit, since they minimize an L_2 functional without any bias with respect to the location of input values. The Legendre polynomial with a multi-index $\boldsymbol{p} = (p_1, p_2, \ldots, p_d)$ is a multivariate polynomial function of d variables $(\eta_1, \eta_2, \ldots, \eta_d) = \boldsymbol{\eta}$ defined by

$$\Psi_{\boldsymbol{p}}(\boldsymbol{\eta}) = \psi_{p_1}(\eta_1)\psi_{p_2}(\eta_2)\cdots\psi_{p_d}(\eta_d), \tag{7}$$

where $\psi_{p_i}(\eta)$ is the standard 1D Legendre polynomial of degree p_i, for $i = 1, 2, \ldots, d$. By convention, the sum of all degrees $p_1 + p_2 + \cdots + p_d$ is called the order of the multivariate Legendre polynomial $\Psi_{\boldsymbol{p}}(\boldsymbol{\eta})$.

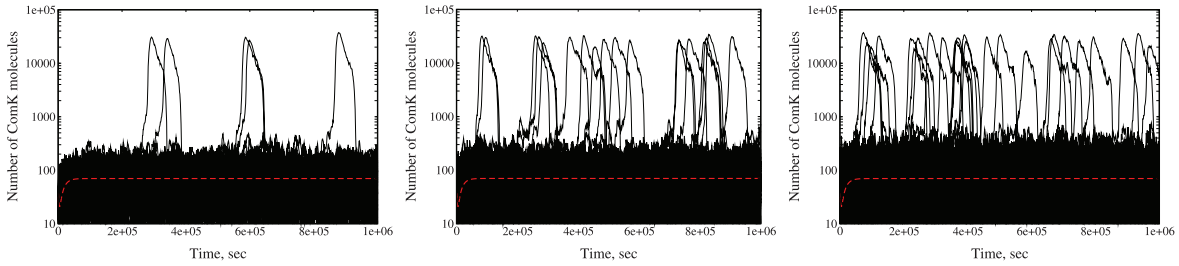

Fig. 3. Three different sets of runs of the stochastic system sharing the same ODE limit (dashed line). The left plot corresponds to the smaller noise level, while the middle plot is the nominal run, and the right plot illustrates sample realizations with the highest noise level.

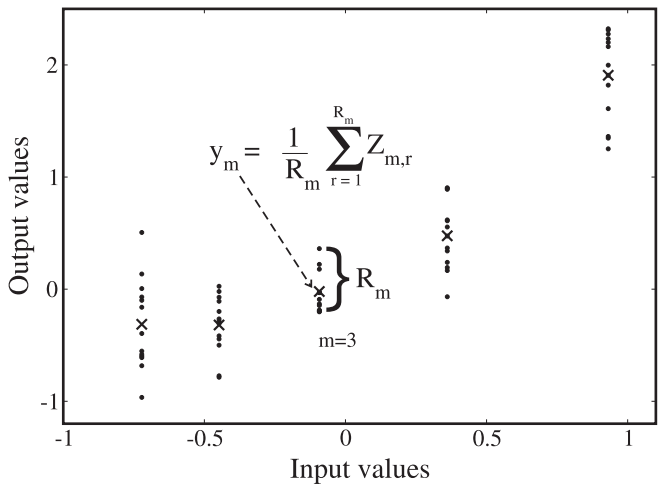

Fig. 4. Illustration of the available training data set. In this example, the number of distinct input parameters is set to $M = 5$. At each parameter location there are $R_m = R = 13$ replica simulations, with the averages marked by a cross symbol.

The input parameter vector $\boldsymbol{\lambda}$ is related to the Uniform [-1,1] random variable vector $\boldsymbol{\eta}$ by the inverse cumulative distribution function (CDF) $F_{\lambda_i}(\cdot)$ of each input parameter (given that they are independent by construction)

$$\lambda_i = F_{\lambda_i}^{-1}\left(\frac{\eta_i + 1}{2}\right), \qquad \text{for } i = 1, 2, \ldots, d. \tag{8}$$

For example, if λ_i is assumed uniform on $[a_i, b_i]$, then

$$\lambda_i = \frac{a_i + b_i}{2} + \frac{b_i - a_i}{2}\ \eta_i. \tag{9}$$

Note that when input parameters are dependent, one can employ the *Rosenblatt* transformation [22] to map the dependent input parameter vector $\boldsymbol{\lambda}$ to a vector of i.i.d. uniform random variables $\boldsymbol{\eta}$. The transformation is essentially a multidimensional generalization of the CDF transform (8) and has been applied in a similar context to a climate land model [23].

Given the map from (8), which is denoted by $\boldsymbol{\lambda} = g(\boldsymbol{\eta})$, the dependence of the output P_c on the input parameter vector $\boldsymbol{\lambda}$ is encapsulated in a polynomial chaos expansion

$$P_c(\boldsymbol{\eta}) \approx y_{\boldsymbol{c}}(\boldsymbol{\eta}) \equiv \sum_{k=0}^{K} c_k \Psi_k(\boldsymbol{\eta}), \tag{10}$$

where the scalar subscript k corresponds to the graded lexicographic ordering of the multi-indices $\boldsymbol{p}$ [24]. The truncation of the above series is chosen according to the order of polynomials. The expansion in (10) retains only polynomials of order up to l, i.e., $p_1 + p_2 + \cdots + p_d \le l$, where $K + 1 = (d + l)!/(d!l!)$ is the total number of terms.

The problem of building the response surface that represents the input-output relationship now reads as follows: given *training* runs at M parameter locations, with R_m replica simulations at each parameter location $\boldsymbol{\lambda}_m = g(\boldsymbol{\eta}_m)$ for $m = 1, \ldots, M$, find the coefficients or *PC modes* $\boldsymbol{c} = (c_0, c_1, \ldots, c_K)$ of the PC expansion in (10). See Fig. 4 for illustration of the available training runs, where $Z = P_c$ is the output observable.

Whenever necessary, we will use the tilde notation $\tilde{\boldsymbol{\lambda}} = g(\tilde{\boldsymbol{\eta}}_n)$ for $n = 1, \ldots, N$, where $N = \sum_{m=1}^{M} R_m$, to reindex the parameter locations that are not necessarily distinct. Also, for clarity of presentation, in all the test cases discussed in this paper, the number of replicas is kept the same for all parameter settings, i.e., $R_m = R$ for $m = 1, 2, \ldots, M$, although this is not required.

For each parameter location, the best guess of the observable would be the data average $y_m = \frac{1}{R_m}\sum_{r=1}^{R_m} Z_{m,r}$. The averages y_m will still be away from the true value $P_c(\boldsymbol{\lambda}_m)$ due to finite sampling. By the Central Limit Theorem (CLT), the data averages y_m will be normally distributed with a variance that is smaller than that of the raw data $Z_{m,r}$ by a factor of $\sqrt{R_m}$.

Orthogonal projection approaches that are based on quadrature integration can efficiently find the PC coefficients $\boldsymbol{c}$ by making use of the polynomial bases orthogonality. In high dimensions, one has to implement *sparse* quadrature integration in order to alleviate the so-called curse of dimensionality [25]. However, while well suited for smooth and deterministic functions, these methods are infeasible in the stochastic context, since the sparse quadrature integration suffers from a noise amplification due to negative weights. Instead, we employ a Bayesian approach that provides a probabilistic answer with any number of training simulations located arbitrarily in the parametric space. It works well with intrinsically stochastic systems and leads to an uncertain response surface representation, the uncertainty being associated with lack-of-knowledge, i.e., with the finite sampling. While in general Bayesian methods are computationally demanding, we employ conjugate prior and Gaussian likelihood constructions in order to arrive to an analytically tractable joint posterior distribution for the PC modes. The next section details this Bayesian inference methodology.

4 Bayesian Inference of Polynomial Chaos Coefficients

Bayesian methods are well suited for dealing with uncertainties from different sources, from intrinsic noise in the system to parametric uncertainty and experimental errors, see [18], for example. Here, we will outline and argue for a Bayesian inference approach to obtain PC representation in (10) with uncertainties *given* a number of forward model runs—training runs—at arbitrarily distributed parameter locations. The approach relies on Bayes' formula

$$P(\mathcal{M}|\mathcal{D}) = \frac{P(\mathcal{D}|\mathcal{M})P(\mathcal{M})}{P(\mathcal{D})}, \tag{11}$$

where $\mathcal{M}$ is the PC model, i.e., it is identified with the polynomial $y_{\boldsymbol{c}}(\cdot)$ or the coefficient vector $\boldsymbol{c}$, while $\mathcal{D}$ is the available data or the available model simulation results. The *prior* probability $P(\mathcal{M})$ and the *posterior* probability $P(\mathcal{M}|\mathcal{D})$ represent degrees of belief about a model $\mathcal{M}$ before and after the data are available, respectively. The *evidence* $P(\mathcal{D})$ does not depend on $\mathcal{M}$ and is simply a normalizing constant if one is only interested in inferring the model parameters. The evidence becomes important in model comparison studies, described in Section 5.2. The key

concept in Bayes' formula is the *likelihood* of the data set, viewed as a function of the model, $P(\mathcal{D}|\mathcal{M}) = L(\mathcal{M})$. Let us rewrite Bayes' formula in the present context of the PC model from (10)

$$q(\boldsymbol{c}) \propto L_{\mathcal{D}}(\boldsymbol{c})p(\boldsymbol{c}), \tag{12}$$

where $p(\cdot)$ and $q(\cdot)$ are prior and posterior multivariate PDFs for the PC coefficients, respectively. The likelihood function $L_{\mathcal{D}}(\boldsymbol{c})$ represents the likelihood of the data given the model, expressed in terms of the discrepancy between the training data (available model simulation results) and the PC model.

The mean of the data $Z_{m,r}$ and its variance at the mth parameter location are, respectively,

$$y_m = \frac{1}{R_m}\sum_{r=1}^{R_m} Z_{m,r}, \tag{13}$$

and

$$s_m^2 = \frac{1}{R_m - 1}\sum_{r=1}^{R_m}(Z_{m,r} - y_m)^2. \tag{14}$$

The Central Limit Theorem suggests a Gaussian distribution, with variance $E = s_m^2/R_m$, for the distance between the true mean and the data mean y_m. As such, the likelihood can be written as:

$$L_{\mathcal{D}}(\boldsymbol{c};\boldsymbol{v}) = \frac{1}{(2\pi)^{M/2}\prod_{m=1}^{M}\sigma_m}\exp\left(-\sum_{m=1}^{M}\frac{\left(y_m - y_{\boldsymbol{c}}(\boldsymbol{\eta}_m)\right)^2}{2\sigma_m^2}\right), \tag{15}$$

where $\boldsymbol{v} = (\sigma_1^2, \ldots, \sigma_m^2)$ are the variances at each parameter location. It is a parameter vector of the likelihood and will be inferred as *hyperparameter* vector along with the main object of inference, the vector of PC coefficients $\boldsymbol{c}$. The vector of variances $\boldsymbol{v}$ is equipped with its own prior reflecting our knowledge $\boldsymbol{v}_m \approx s_m^2/R_m$, and, consequently, we replace the simple Bayesian formulation in (12) with a more general one

$$\tilde{q}(\boldsymbol{c},\boldsymbol{v}) \propto L_{\mathcal{D}}(\boldsymbol{c};\boldsymbol{v})p(\boldsymbol{c})\hat{p}(\boldsymbol{v}), \tag{16}$$

where the joint posterior $\tilde{q}(\boldsymbol{c},\boldsymbol{v})$ will subsequently be marginalized over the hyperparameter vector $\boldsymbol{v}$ to obtain the posterior $q(\boldsymbol{c})$ for the object of inference, i.e., model parameter vector $\boldsymbol{c}$:

$$q(\boldsymbol{c}) = \int_0^\infty \tilde{q}(\boldsymbol{c},\boldsymbol{v})d\boldsymbol{v}. \tag{17}$$

For general likelihood functions and prior distributions, one obtains the posterior distribution by actually sampling from it using MCMC techniques [26], [27]. However, with a Gaussian likelihood in (15) and appropriate prior distributions there is an analytic form for the posterior distribution. With no information on values of PC coefficients, it is reasonable to assume independent uniform priors on each coefficient c_k (with infinite domain, i.e., this prior is improper), while the inverse-gamma family of distributions $\mathcal{IG}(\alpha_m, \beta_m)$ offers sufficient flexibility for positive-valued distributions, at the same time allowing for analytical tractability, and reads as

$$\hat{p}(\boldsymbol{v}) = \prod_{m=1}^{M}\hat{p}(\sigma_m^2) = \prod_{m=1}^{M}\frac{\beta_m^{\alpha_m}}{\Gamma(\alpha_m)}(\sigma_m^2)^{-\alpha_m-1}e^{-\frac{\beta_m}{\sigma_m^2}}, \tag{18}$$

one obtains the joint posterior for $(\boldsymbol{c},\boldsymbol{v})$:

$$\tilde{q}(\boldsymbol{c},\boldsymbol{v}) \propto \prod_{m=1}^{M}\frac{1}{\sigma_m}\exp\left(-\frac{\left(y_m - y_{\boldsymbol{c}}(\boldsymbol{\eta}_m)\right)^2}{2\sigma_m^2}\right) \times \frac{\beta_m^{\alpha_m}}{\Gamma(\alpha_m)}(\sigma_m^2)^{-\alpha_m-1}e^{-\frac{\beta_m}{\sigma_m^2}}. \tag{19}$$

Marginalizing over the hyperparameter vector $\boldsymbol{v}$ leads to the following marginal distribution for the PC coefficient vector $\boldsymbol{c}$:

$$q(\boldsymbol{c}) \propto \prod_{m=1}^{M}\frac{\Gamma(\alpha_m + 1/2)}{\Gamma(\alpha_m)}\beta^{-1/2}\left(1 + \frac{\left(y_m - y_{\boldsymbol{c}}(\boldsymbol{\eta}_m)\right)^2}{2\beta_m}\right)^{\alpha_m+1/2}. \tag{20}$$

In the following, we describe how to select values for (α_m, β_m). The available data set $\mathcal{D}$ carries very little information on the values of variances σ_m^2, rendering the inference of $\boldsymbol{v}$ challenging. At the mth parameter location one can compute the variance estimate s_m^2 from (14). However, it is the only information we have about the true value σ_m^2. Hence, a certain degree of regularization is needed. Namely, carefully picked prior parameters (α_m, β_m) will restrict the range of σ_m^2 for each $m = 1, 2, \ldots, M$. The following informal arguments help in the selection of parameters α_m and β_m that are consistent with the single known value s_m^2 of the sample variance of R_m replicas. The coefficient of variation, i.e., the ratio of the standard deviation to the mean, of a $\mathcal{IG}(\alpha_m, \beta_m)$ random variable is $1/\sqrt{\alpha_m - 2}$, while the coefficient of variation of a sample variance, viewed as a random variable, is $2/\sqrt{R_m - 1}$, for normal random variables. This suggests a rule-of-thumb choice of α_m:

$$\alpha_m = \frac{R_m + 3}{2}. \tag{21}$$

Furthermore, the most likely value for a single sample from an $\mathcal{IG}(\alpha_m, \beta_m)$ distribution is the mode of that distribution $\beta_m/(\alpha_m + 1)$. Therefore, a reasonable choice for β_m, given a single sample variance estimate for each m, s_m^2, would be

$$\frac{\beta_m}{\alpha_m + 1} = \frac{s_m^2}{R_m} \quad \text{due to CLT, and, given (21),}$$
$$\beta_m = s_m^2\frac{R_m + 5}{2R_m}. \tag{22}$$

As an illustration, Fig. 5 demonstrates several prior distributions for $s_m^2 = 0.1$ and various R_m. Clearly, for large values of R_m the prior for σ_m^2 is quite narrow, the limiting case $R_m \to \infty$ being equivalent to fixing the hyperparameters to values derived from the sample variance, i.e., s_m^2/R_m. In this case, we have the simple Bayesian formulation (12) with a likelihood

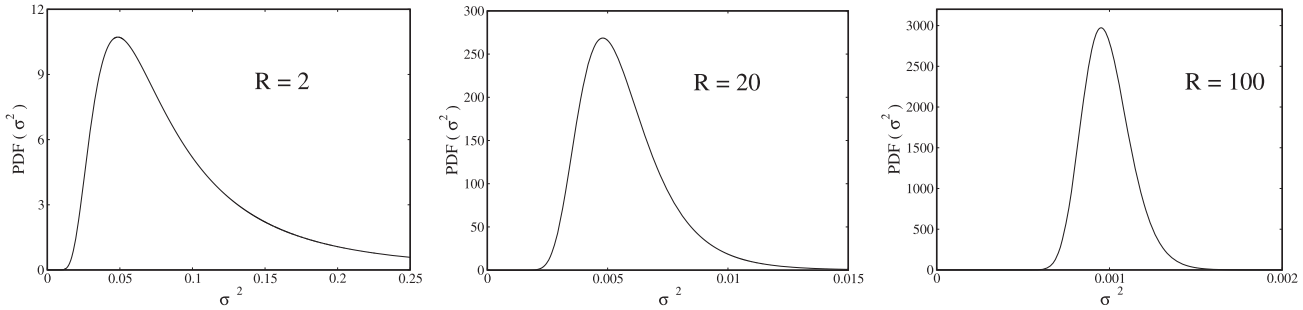

Fig. 5. Demonstration of inverse gamma probability distributions with parameters α and β picked according to a given replica number, R, and $s^2 = 0.1$, using (21) and (22). The subscript m is dropped for clarity of presentation.

$$L_{\mathcal{D}}(\boldsymbol{c}) = L_{\mathcal{D}}(\boldsymbol{c}; \boldsymbol{s}^2) = \frac{1}{(2\pi)^{M/2} \prod_{m=1}^{M} (s_m/\sqrt{R_m})} \exp\left(- \sum_{m=1}^{M} \frac{(y_m - y_{\boldsymbol{c}}(\boldsymbol{\eta}_m))^2}{2s_m^2/R_m} \right), \tag{23}$$

and uniform priors on each c_k. This formulation leads to a multivariate normal posterior distribution

$$\boldsymbol{c} \in \mathcal{MVN}(\underbrace{(\boldsymbol{\Psi}^T \boldsymbol{Q}^{-1} \boldsymbol{\Psi})^{-1} \boldsymbol{\Psi}^T \boldsymbol{Q}^{-1} \boldsymbol{y}}_{\boldsymbol{\mu}}, \underbrace{(\boldsymbol{\Psi}^T \boldsymbol{Q}^{-1} \boldsymbol{\Psi})^{-1}}_{\Sigma}), \tag{24}$$

where $\boldsymbol{\Psi}$ is an $M \times (K+1)$ matrix with elements $\boldsymbol{\Psi}_{mk} = \Psi_k(\boldsymbol{\eta}_m)$ and $\boldsymbol{Q}$ is a diagonal matrix with entries $Q_{mm'} = \delta_{m,m'} R_m / (2s_m^2)$. The maximum a posteriori (MAP) value for the object of inference $\boldsymbol{c}$ coincides with the maximum likelihood estimate (MLE), since the priors are uniform,

$$\boldsymbol{\mu}_{\boldsymbol{c}} = (\boldsymbol{\Psi}^T \boldsymbol{Q}^{-1} \boldsymbol{\Psi})^{-1} \boldsymbol{\Psi}^T \boldsymbol{Q}^{-1} \boldsymbol{y}, \tag{25}$$

and it is also the solution of a weighted least squares problem

$$\underset{\boldsymbol{c}}{\operatorname{argmin}} \sum_{m=1}^{M} \frac{(y_m - y_{\boldsymbol{c}}(\boldsymbol{\eta}_m))^2}{2s_m^2/R_m}. \tag{26}$$

Note that it can be shown that the solution in (20) of the hierarchical Bayesian formulation from (16) converges to the multivariate normal posterior distribution outlined in (24) in the limit $R_m \to \infty$ and (α_m, β_m) picked according to (21) and (22).

Furthermore, having obtained a probabilistic representation (24) of PC coefficients $\boldsymbol{c}$, one can conclude that the output representation

$$y_{\boldsymbol{c}}(\boldsymbol{\eta}) = \sum_{k=0}^{K} c_k \Psi_k(\boldsymbol{\eta}), \tag{27}$$

is a Gaussian process with mean function

$$m(\boldsymbol{\eta}) = \sum_{k=0}^{K} \boldsymbol{\mu}_k \Psi_k(\boldsymbol{\eta}), \tag{28}$$

and covariance

$$C(\boldsymbol{\eta}, \boldsymbol{\eta}') = \sum_{i,j=0}^{K} \Psi_i(\boldsymbol{\eta}) \Sigma_{ij} \Psi_j(\boldsymbol{\eta}'). \tag{29}$$

In other words, our final representation is an *uncertain* response surface, the uncertainty being a direct consequence of the posterior distribution of the PC modes. The more training samples are taken, the narrower the posterior distribution is, since it is associated with the lack-of-knowledge in the Bayesian framework.

We note that one can build multiple response surfaces for multiple output quantities of interest, as long as these quantities are evaluated at the training points. One also can lump a family of outputs into a single one, if they are related through a parameter. For example, the competence transition threshold value of $C_{\text{thr}} = 5{,}000$ is somewhat arbitrary. To study the influence of the choice of this threshold, one could treat C_{thr} as an additional input parameter to construct the response surface over, thereby enabling the analysis of the output quantities with respect to the chosen threshold value.

5 Challenges and Postprocessing

In this section, we will briefly discuss some of the general issues with the PC representation in (10) and specific challenges associated with the SRN of competence in *B.subtilis*.

5.1 Choice of the Input Points and PC Order

The multivariate polynomial in (10) has $K+1 = (d+l)!/(d!l!)$ terms, where l is the truncation order. Typically, the Bayesian solution formulated in Section 4 exists only if the number of distinct input points is larger than the degrees of freedom in the polynomial representation, i.e., $M > K+1$. For a given number of input points, the choice of an appropriate order is crucial, because with only a limited number of training evaluations, one needs to avoid *overfitting*, when M is too close to $K+1$. In this work, we will simply set aside a set of input values for validation tests. Specifically, for high-dimensional problems, these validation points are chosen randomly according to a latin hypercube sampling (LHS) to fill in the space reasonably well. The outcome of the validation tests will serve as an error measure of the particular polynomial representation. Ideally, one could infer polynomial expansions of increasing orders to a given data, and then choose an order that is near the "elbow" of the error-versus-order graph, i.e., where the error starts to decrease at a much slower rate. Alternatively, the choice of the order can be driven by a given threshold—whenever the error is not small enough, the order of the representation should be increased.

5.2 Error Measure

A commonly used error measure is the relative L_2 error between the true function and its representation, i.e.,

$$\frac{\|P_c(\boldsymbol{\eta}) - y_{\boldsymbol{c}}(\boldsymbol{\eta})\|_2}{\|P_c(\boldsymbol{\eta})\|_2}. \tag{30}$$

However, the norm in the numerator is difficult to compute with quadrature integration: it either suffers from the curse of dimensionality (full tensor product) or is extremely inaccurate due to the noise in computing estimates of $P_c(\boldsymbol{\eta})$ (sparse quadrature). Therefore, we estimate the norms via LHS. Specifically, denote by $V = \{\boldsymbol{\zeta}_i\}_{i=1}^{N_V}$ a set of randomly chosen input points obtained via LHS. The error measure of a particular representation

$$P_c(\boldsymbol{\eta}) \approx y_{\boldsymbol{c}}(\boldsymbol{\eta}) \equiv \sum_{k=0}^{K} c_k \Psi_k(\boldsymbol{\eta}), \tag{31}$$

is taken to be

$$E(\boldsymbol{c}; V) = \sqrt{\frac{\sum_{i=1}^{N_V} (P_c(\boldsymbol{\zeta}_i) - y_{\boldsymbol{c}}(\boldsymbol{\zeta}_i))^2}{\sum_{i=1}^{N_V} P_c(\boldsymbol{\zeta}_i)^2}}, \tag{32}$$

where P_c is the competence probability estimate computed according to (6). In other words, we take the relative l_2 distance between the actual function evaluations and the representation at given validation points. Note that the cardinality N_V of the validation set V should not be too large for expensive forward functions as $P_c(\boldsymbol{\zeta}_i)$ itself is estimated via replica simulations at an input location $\boldsymbol{\zeta}_i$, for $i = 1, 2, \ldots, N_V$. Also, we typically compute the error with two or three different validation sets to gauge confidence in the estimate and ensure that N_V is sufficiently large to produce a reliable estimate. Using replica validation sets is particularly important, since the accuracy of the error measure is also limited by the leftover noise in function evaluations $P_c(\boldsymbol{\zeta}_i)$.

5.3 Constrained Output Values: Mapped PC Expansion

The probability of competence in *B.subtilis* is restricted to the range $[0, 1]$. Due to the oscillatory nature of polynomials, such a physical limit is hard to enforce with PC expansions. In such cases, a transformation to a real line ensures that the polynomial-based representation does not leave the physical bounds of the observable of interest. To ensure that the representation with a polynomial basis is in the interval $[0, 1]$, we employ a map $H : (-\infty, +\infty) \mapsto [0, 1]$ and its inverse $H^{-1}(\cdot)$ by

$$\begin{aligned} \tilde{y} &= H(y) = \frac{e^{-y}}{e^{-y} + e^{y}}, \qquad \text{and} \\ y &= H^{-1}(\tilde{y}) = \log\left(\frac{\tilde{y}}{1 - \tilde{y}}\right), \end{aligned} \tag{33}$$

such that $\tilde{y}$ is guaranteed to be in $[0, 1]$, regardless of the range in y. The final representation now takes the form

$$\hat{y} = H\left(\sum_{k=0}^{K} c_k \Psi_k(\boldsymbol{\eta})\right). \tag{34}$$

Therefore, effectively we look at $H^{-1}(p)$ as our model output, where p is the probability of competence. Due to this transformation, the global sensitivity indices computation is not as straightforward as for a plain polynomial expansion, but Monte-Carlo sampling still allows estimating these indices.

5.4 Nonsmooth Output Behavior: Piecewise or Mixed PC Expansions

While exploring the input parameter space, the output P_c varies between a fully vegetative ($P_c = 0$) and fully competent ($P_c = 1$) state, often with a sharp transition in between. Generally, such nonlinearities are challenging polynomial surrogates. Typically, input domain decomposition techniques serve as a remedy against discontinuous or strongly nonlinear output behavior [28], [29], [30]. However, with only limited number of sample runs available in high dimensions, domain decomposition is not straightforward. Naive domain splitting algorithms do not scale well to high dimensions. Below we propose a clustering-based classification approach that classifies input points according to their corresponding outputs and then generates a piecewise PC construction. Such clustering techniques are generally introduced in [31], where the authors developed methods for input clustering that take into account output values to enhance functional representations. In our case, clustering will be purely based on the output values, in order to separate out the plateaus $P_c = 0$ and $P_c = 1$ in the forward function. In general, assume that the input sample point set $\mathcal{S}$ is divided into L clusters $\mathcal{S}_1, \ldots, \mathcal{S}_L$ according to the criterion based on the corresponding output values. In each cluster, one has a global PC (or mapped PC) expansion

$$P_c^{(l)}(\boldsymbol{\eta}) = \sum_{k=0}^{K} c_k^{(l)} \Psi_k(\boldsymbol{\eta}), \quad \text{for } l = 1, \ldots, L. \tag{35}$$

In particular, for the representation of the probability of competence, the input parameter space is naturally divided into three regions: fully vegetative ($P_c = 0$), fully competent ($P_c = 1$) or neither of those ($0 < P_c < 1$). The first two regions, clearly, are represented by a constant, or zeroth order PC, while the more meaningful, transition region is represented by PCs of increasing orders until the error is satisfactory. Since the two other PCs are trivial, we will refer to the order of this "middle" PC as the order of the expansion.

In predictive mode, for an unknown $\boldsymbol{\eta}$, we combine these expansions according to a nearest neighbor search in the input space:

$$P_c(\boldsymbol{\eta}) = P_c^{(l)}(\boldsymbol{\eta}), \quad \text{if } \mathrm{nn}_1(\boldsymbol{\eta}) \in \mathcal{S}_l, \tag{36}$$

where $\mathrm{nn}_1(\boldsymbol{\eta})$ denotes the (first) nearest neighbor of $\boldsymbol{\eta}$ according to the Euclidian distance in the input space.

For a representation that is smoother if sufficiently many training runs are performed, one can generalize this piecewise PC construction by taking multiple nearest neighbors into account to obtain a *mixture* of appropriately weighted PC expansions in the following way:

$$P_c(\boldsymbol{\eta}) = \sum_{i=1}^{k} w^{(i)}(\boldsymbol{\eta}) P_c^{(l(i))}(\boldsymbol{\eta}), \tag{37}$$

where $l(i)$ is the cluster corresponding to the ith nearest neighbor of $\boldsymbol{\eta}$ and the weights are chosen according to the inverse distance

$$w^{(i)}(\boldsymbol{\eta}) = \frac{d^{-1}(\boldsymbol{\eta}; i)}{\sum_{i=1}^{k} d^{-1}(\boldsymbol{\eta}; i)}, \tag{38}$$

i.e., $d^{-1}(\boldsymbol{\eta}; i)$ is the inverse of the distance from $\boldsymbol{\eta}$ to its ith nearest neighbor. It can be verified that for $k = 1$ case, the mixed PC expansion in (37) reduces to the piecewise PC formulated in (35).

5.5 Variance-Based Sensitivity Indices

After the PC representation in (10) is built with respect to a range of input parameter variations, one can extract sensitivity information according to a variance decomposition [32]. The main effect sensitivity indices S_i are defined as

$$S_i = \frac{\mathrm{Var}[\mathbb{E}(y_{\boldsymbol{c}}(\boldsymbol{\eta})|\eta_i)]}{\mathrm{Var}[y_{\boldsymbol{c}}(\boldsymbol{\eta})]}, \tag{39}$$

for $i = 1, \ldots, d$, while the joint sensitivity indices S_{ij} are

$$S_{ij} = \frac{\mathrm{Var}[\mathbb{E}(y_{\boldsymbol{c}}(\boldsymbol{\eta})|\eta_i, \eta_j)]}{\mathrm{Var}[y_{\boldsymbol{c}}(\boldsymbol{\eta})]} - S_i - S_j, \tag{40}$$

for $i, j = 1, \ldots, d$. These indices are also called Sobol indices [33]. The variances in the numerators of (39) and (40) are with respect to the fixed variables η_i or (η_i, η_j), while the expectations are with respect to the rest of the variables. The sensitivity index S_i can be interpreted as the fraction of the variance in the output that can be attributed to the ith input only. Similarly, S_{ij} is the variance fraction that is due to the joint contribution of ith and jth inputs only. One can similarly define joint sensitivity indices for multivariate couplings as well. By definition, all the indices sum up to one, enabling fair comparison between main effect and joint sensitivity indices. Although simple analytical expressions for S_i and S_{ij} are available in terms of PC coefficients c_k, we will estimate these indices via Monte-Carlo sampling, in light of the adjustments to the PC representation outlined in Section 5.4.

6 Results and Discussion

In the 2D analysis, described below, we will vary the constitutive transcription rates for comK and comS, respectively, i.e., parameters k_1 and k_4, since preliminary studies as well as previous work by Suel et al. [13] suggested that these are very important parameters. In the subsequent 18D analysis the input parameter vector will contain almost all reaction rate parameters (all except n and p, see Table 1). In order to enforce positivity, the logarithms of rate constants are considered as input rather than the rate constants themselves.

6.1 2D Analysis

Let us consider a 2D case to illustrate the uncertain response surface construction, outlined in Section 3. The two input parameters for uncertainty analysis are chosen to be the rate constants k_1 and k_4. These are the constants analyzed in [13]. First, to illustrate the dependence of the competence dynamics on the selected input parameters, we show the contour plot of the dependence of the competence fraction on the two selected input parameters; see Fig. 6. The input parameters are perturbed in the *logarithmic* scale, by a factor of $f = 5$. That is, the input domain of interest is $[\frac{1}{f}\tilde{k}_1, f\tilde{k}_1] \times [\frac{1}{f}\tilde{k}_4, f\tilde{k}_4]$, where the nominal values are denoted with a tilde. In the logarithmic scale, clearly, this corresponds to having the nominal values in the center of the domains, i.e., $\lambda_1 = \log k_1 = \log \tilde{k}_1 + \eta_1 \log f$ and $\lambda_2 = \log k_4 = \log \tilde{k}_4 + \eta_2 \log f$ for $\eta_1,\ \eta_2 \in [-1, 1]$. Note that the contour plots indicate that large portions of the input domain correspond to purely vegetative ($P_c = 0$) or purely competent ($P_c = 1$) regimes. Also, the output P_c is physically constrained to be in $[0, 1]$. Therefore, we apply the mapping and clustering techniques described in Section 5.4. Fig. 7 demonstrates the maximum a posteriori response surface fit resulting from a fourth order expansion of the mapped output enhanced with nearest neighbor clustering-based polynomial mixture methodology, which clusters out values that are exactly 0 and 1, i.e., it represents them with a constant "polynomial" fit. The number of training parameter locations is set to $M = 200$, while at each location we simulated $R = 100$ replicas of the model to obtain an estimate of P_c.

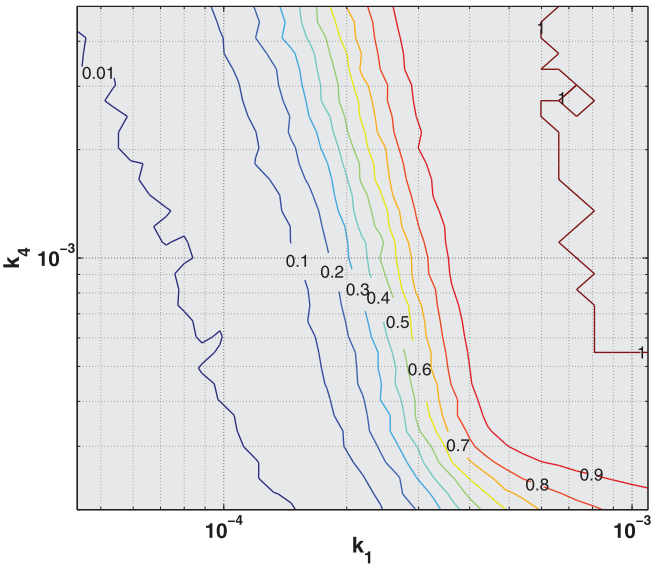

Fig. 6. Demonstration of the output dependence on the two selected input parameters. The results are based on $M = 200$ sampled parameter locations with $R = 100$ replica simulations at each parameter value.

Fig. 8 demonstrates the relative l_2 error reduction, with respect to a number of training points for a fourth order polynomial case, and with respect to the polynomial order for a fixed number of training points $M = 200$, respectively. The details of the error computation are given in Section 5.2. Since the error is computed relying on a Monte-Carlo set of 1,000 validation samples with $R = 100$ replica SSA simulations per sample, we replicated the error computation three times to have some confidence in the estimate. The resulting 5 percent relative error is approximately of the order of the

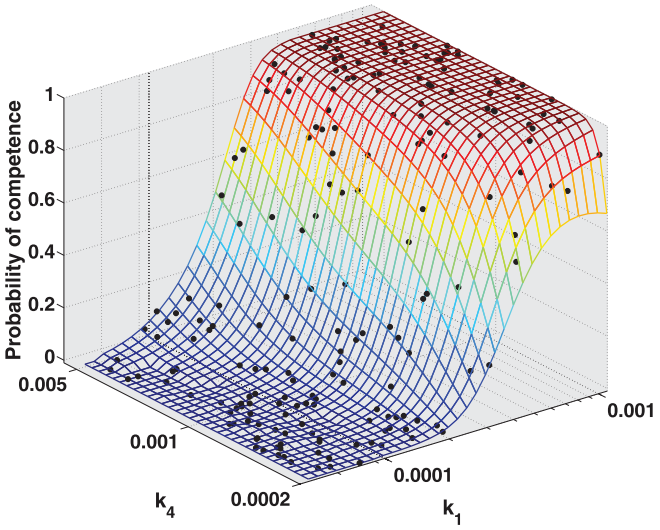

Fig. 7. Fourth order representation, together with the training points. The results are based on $M = 200$ sampled parameter locations with $R = 100$ replica simulations at each parameter value.

data noise, and cannot be improved with higher order or higher number of training points. We have also found that the uncertainty associated with the joint posterior distribution, on the other hand, is much smaller than the data noise, suggesting that $M = 200$ sample points is sufficient for determining the PC modes of the current representation.

6.2 High-Dimensional Analysis and Dimensionality Reduction

For a complete parametric uncertainty study, we now allow perturbations in essentially all input parameters. Namely, all parameters are considered from Table 1, except for $n = 2$ and $p = 5$; a total of 18 input parameters. While a mixture PC expansion can be used to properly represent the different regimes in P_c, preliminary studies have shown that the accuracy of mixture PC expansions is challenged by misclassification due to sparsity of available data in high dimensions. Typically, three-five points per dimension are required for the single nearest neighbor algorithm to produce reliable results with reasonably small errors. In the 18D case, this would require at least 3^{18} training points. Therefore, we perform a dimensionality reduction based on sensitivity indices first, and then perturb only those input parameters that have a strong effect on the output of interest. Specifically, we obtain a *global* PC expansion for the mapped variable $\log\frac{P_c}{1-P_c}$. Although the expansion is not very accurate near the "plateau" regions of $P_c = 0$ or $P_c = 1$, it still provides an approximation that can inform us about the importance of various parameters. We then compute sensitivity indices S_i from (39) for all dimensions $i = 1, \ldots, d$ based on that PC expansion. The sensitivity indices in (39) are computed using Monte-Carlo sampling. To estimate the variance in the numerator of (39) we used 100 samples η_i, and, for each of these values, another 100 samples are drawn to evaluate the variance. The denominator is computed with a 1,000 sample variance estimator. Again, since only a down-selection of parameters is sought, a qualitative differentiation between parameters is sufficient, hence the number of samples picked above leads to a reliable down-selection. The close agreement between two replica Monte-Carlo computations, shown in Fig. 9 indicates that the set $(k_1, k_2, k_k, k_8, k_3, k_7)$ can be reliably selected as the most influential input parameter set. Note that, per Table 1, the selected parameters are all related to the key reactions involving comK expression, confirming the importance of comK dynamics in the transition to competence, and consistent with what could be expected based on biological insight. We have also performed local sensitivity analysis around the nominal parameter set, one parameter at a time. This single-parameter local sensitivity analysis leads to qualitatively similar results, although it is based on local parameter perturbations only.

After selecting the six input parameters that are the most influential on the probability of competence, we apply the nearest neighbor classification to obtain a PC expansion with two constant plateaus in regions with trivial structure (i.e., $P_c = 0$ or $P_c = 1$). The input parameters are perturbed in the 6D space on a logarithmic scale with a factor of $f = 2$ higher or lower than the nominal values. Fig. 10 shows a 2D slice of the resulting 6D mean response surface. Global PC expansions, with or without the output map $[0, 1] \to (-\infty, \infty)$, would have struggled to represent the constant plateaus

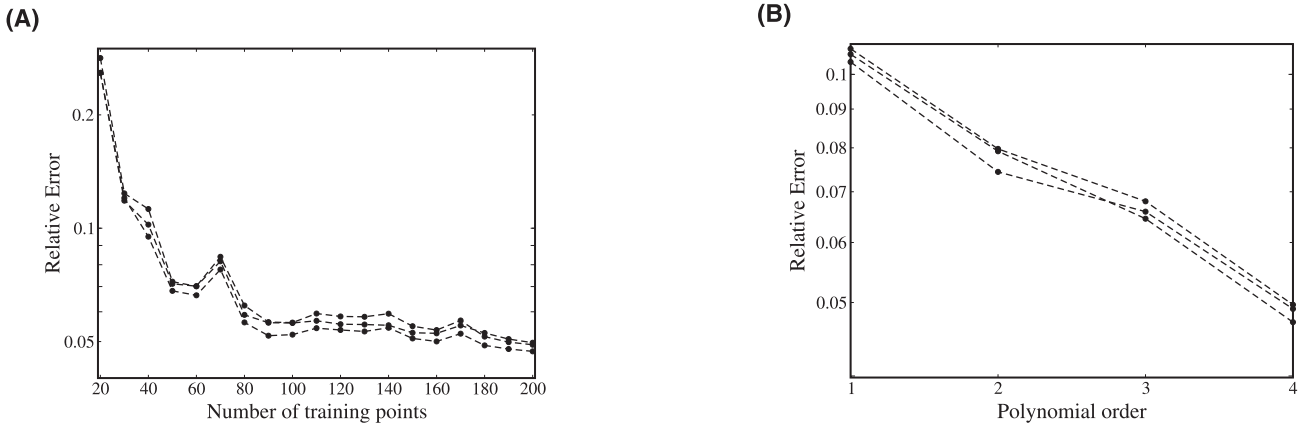

Fig. 8. The error reduction with respect to (A) the number of training points and (B) the order of the underlying polynomial expansion.

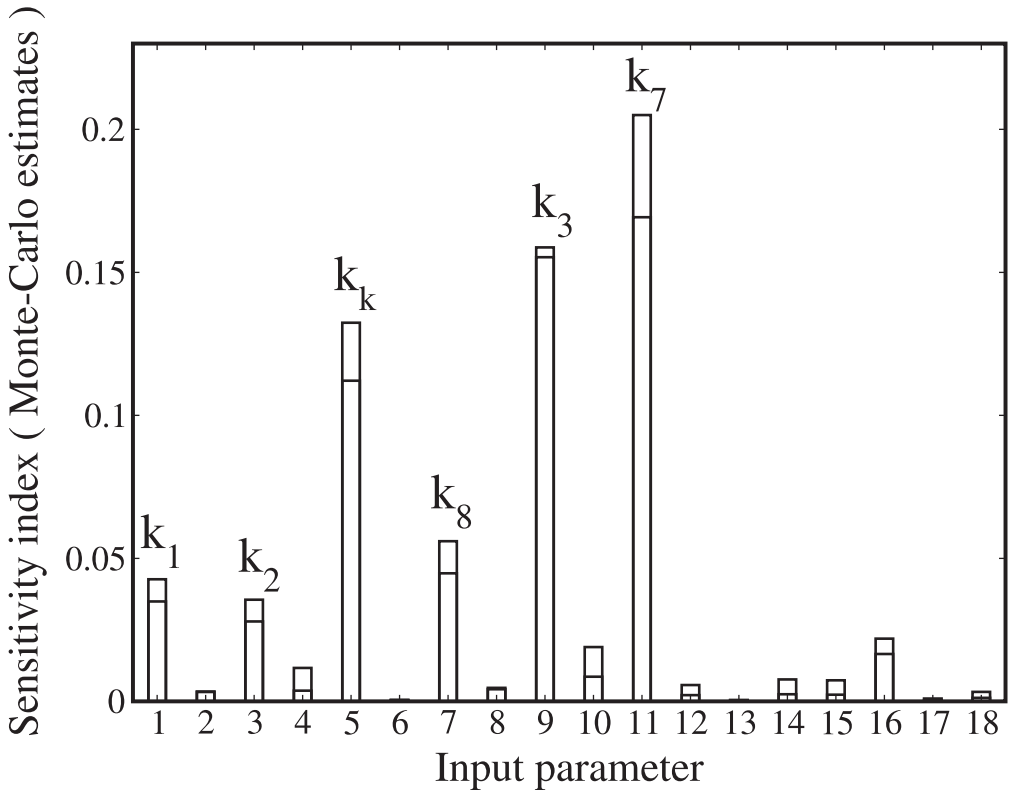

Fig. 9. Sensitivity indices for all the parameters. Two bars are overlaid for each parameter, corresponding to two replica sets of Monte-Carlo samples.

well, while this mixture PC representation "learns" the regions with constant values and improves the accuracy of the resulting response surface. Note that, while we show only the mean expansion, the joint posterior distribution for the parameters of the representation is available. The replica-averaged data noise level ranges from 0 near the plateaus to approximately 0.05 in regions of sharp gradients. In this case too, similar to the 2D analysis, the uncertainty in the response surface that corresponds to the posterior distribution of the PC coefficients is found to be much smaller than the remaining noise level in the data, suggesting that $M = 1{,}000$ parameter samples is sufficient for determining the PC coefficients accurately. We have observed significant reduction in the relative error with increasing the order of the expansion up to fourth order, while higher order expansions did not show considerable error reduction. The relative error between the representation and the data is about 0.08 for the fourth order expansion and is approximately on the same magnitude as the noise level of the replica-averaged data (this comparison between relative and absolute error measures is fair since the data itself vary between 0 and 1), indicating that there is no need for further increase in the order of the expansion.

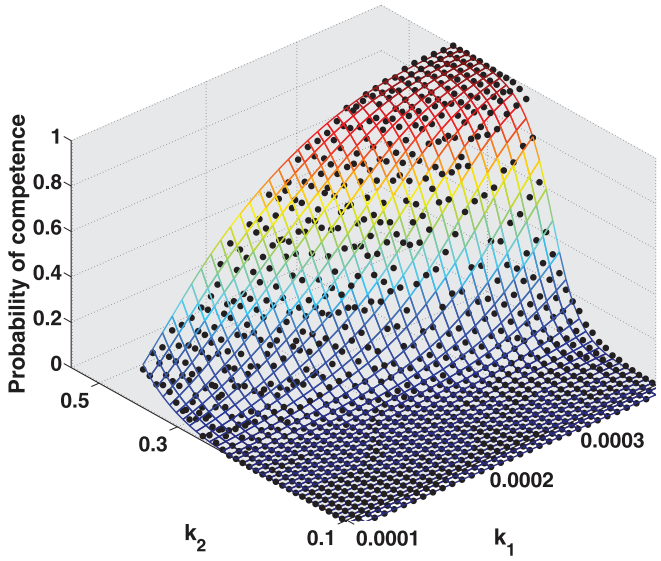

Fig. 10. A 2D slice of a 6D mixture PC expansion. This fourth order mapped, mixture PC surface is inferred using $M = 1{,}000$ training locations with $R = 100$ replica simulations per location. The relative error measure for this response surface is approximately 0.08. The points shown are for visualization only; these are model simulation results on this particular 2D slice and do not coincide or overlap with the set of training points.

Fig. 11 illustrates the main effect and joint sensitivity indices from (39) and (40) by Monte-Carlo sampling of the six input parameters. The main effect indices show a similar ranking of input parameters as in the 18D study from Fig. 9. Beyond that, the inspection of joint sensitivity indices S_{ij} indicates that, for example, the input parameter couples (k_3, k_7) and (k_2, k_7) have a stronger contribution to the output uncertainty than k_1, k_2, and k_8 individually. This suggests that, if one needs to reach a certain output target, it is useful to vary parameters k_3 and k_7 jointly before trying to perturb k_1, k_2 or k_8 individually. Also note that the large-scale dynamics of the system are controlled by the ODE from (1). Some of the ODE parameters in (4) depend explicitly on combinations k_1k_3/k_7 and k_2k_3/k_7, explaining

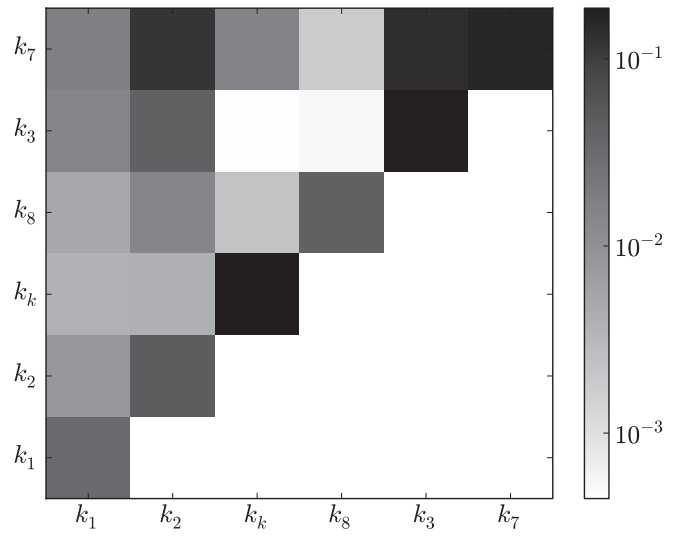

Fig. 11. Global sensitivity indices based on a six-parameter representation. The diagonal entries correspond to the main effect sensitivity indices from (39), while the superdiagonal entries correspond to the joint sensitivity indices from (40). Indices are on a logarithmic grayscale for visual convenience.

the high joint sensitivity indices of parameter pairs involving k_1, k_2, k_3, and k_7.

Note that the sensitivity analysis task is essentially broken down into two steps: first, a response surface PC approximation is constructed based on a number of training runs, and then the response surface is evaluated many times to obtain sensitivity information. To compare the performance of this approach to conventional Monte Carlo sampling, we have implemented a brute-force sensitivity analysis in which the forward function itself is evaluated. However, given the computational cost of a set of replica SSA simulations, we could only sample the function at 10,000 input parameter configurations using 100 processors of a single machine for about three days. Our results have indicated that these 10,000 runs do not lead to reliable estimates of the sensitivity coefficients for the 6D input parameter case. The methodology we present in this work, on the other hand, while introducing an approximation error due to the response surface construction, allows essentially unlimited sampling of the response surface to obtain converged estimates of the sensitivity indices. Namely, we have repeated the computation of the sensitivity indices depicted in Fig. 11 five times and recovered the same result within 5 percent accuracy based on different response surfaces found by five different initial training sets of 1,000 samples. Note that the computational overhead of evaluating the response surface is negligible compared to that of the forward model simulations. Therefore, we have achieved reasonably accurate (within 5 percent) sensitivity information of the response surface approximation—that is within 0.08 relative error itself—with just 1,000 forward model runs, while the brute-force approach did not produce a stable answer even with 10,000 model simulations (more than 30 percent discrepancy between the results from identical studies).

7 Conclusions

To analyze the noise-driven competence dynamics of $B.subtilis$, we have developed and implemented a Bayesian strategy to infer a polynomial representation for an expected value of a stochastic model with respect to input parameters. Specifically, a PC framework is utilized that offers convenient propagation of input uncertainties to outputs of interest. We indicated that orthogonal projection formulas are either inefficient for large numbers of input parameters (full product quadrature) or are extremely inaccurate for noisy model evaluations (sparse grid quadrature). The Bayesian framework, on the contrary, provides a robust probabilistic answer for any number of noisy model evaluations even in a large dimensional input space. The outcome is a representation of the input-output relationship with parameters of this representation being described by a joint posterior distribution. This *uncertain* response surface can further be used as a surrogate model for inverse problems, for optimization, or any other procedure that requires a prohibitively large number of evaluations of the forward model. Furthermore, in this work, besides the curse of dimensionality and the intrinsic noise in the forward model, a further challenge was the strongly nonlinear input-output relationships, which made a global polynomial response surface inaccurate. To deal with this nonlinearity, we have developed a mixture PC expansion that is based on a nearest neighbor classification. Although low-dimensional proof-of-concept results are encouraging, the nearest neighbor classification is not sufficiently accurate for a large number of dimensions due to the sparsity of the data. In such cases, a dimensionality reduction is first performed using estimates of variance-based sensitivity indices. Also, to deal with constrained outputs and to force the polynomial representation to satisfy these constraints, we use a map, and its inverse, from the constrained region to $(-\infty, +\infty)$. As a benchmark problem, we have used the stochastic reaction network of the competence transition in the $B.\,subtilis$ bacterium, where the input parameters are the rate constants and the output is the probability of competence in the steady state of the system. Both 2D and 6D results are demonstrated, the latter being a result of a down-selection of all 18 input parameters using global sensitivity indices. We note that the sensitivity-based down-selection of input parameters is an automated procedure, and it was found to be in agreement with what could be expected based on biological insight.

We note that perhaps other classification approaches can be used to obtain mixture PC representation that are better suited for high dimensional problems. While the application of more advanced classifiers is part of our ongoing work, we believe that the general strategy of classifying input points according to the respective output values is an appropriate building block of generating mixture PC expansions for nonsmooth input-output relationships. We applied the methodology to a stochastic reaction network of competence dynamics in $B.subtilis$ and confirmed and extended previously published results. Typically, in the literature, the sensitivities in the system are computed via single parameter, local perturbations at some nominal values. Our technique, on the other hand, provides rigorous, quantitative sensitivity analysis taking into account all input parameters in a fully coupled fashion. While the proof-of-concept results are demonstrated on the competence dynamics network of $B.subtilis$, the proposed approach is sufficiently general to be applied semiautomatically to analysis of any stochastic system.

ACKNOWLEDGMENTS

This work was supported by the US Department of Energy (DOE) 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 managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the US Department of Energy's National Nuclear Security Administration under contract No. DE-AC04-94AL85000.

REFERENCES

[1] N. van Kampen, *Stochastic Processes in Physics and Chemistry.* Elsevier Science, 1992.
[2] D. Gillespie, *Markov Processes: An Introduction for Physical Scientists.* Academic Press, 1992.
[3] S. Macnamara, K. Burrage, and R. Sidje, "Multiscale Modeling of Chemical Kinetics via the Master Equation," *Multiscale Modeling and Simulation,* vol. 6, no. 4, pp. 1146-1168, 2008.
[4] 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," *J. Chemical Physics,* vol. 129, pp. 95-105, 2008.
[5] B. Munsky and M. Khammash, "The Finite State Projection Approach for the Analysis of Stochastic Noise in Gene Networks," *IEEE Trans. Automatic Control,* vol. 53, no. 1, pp. 201-214, Jan. 2008.
[6] D. Gillespie, "A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions," *J. Computational Physics,* vol. 22, pp. 403-434, 1976.
[7] D. Gillespie, "Exact Stochastic Simulation of Coupled Chemical Reactions," *J. Physical Chemistry,* vol. 81, no. 25, pp. 2340-2361, 1977.
[8] R. Gunawan, Y. Cao, L. Petzold, and F.J. Doyle, "Sensitivity Analysis of Discrete Stochastic Systems," *Biophysical J.,* vol. 88, no. 4, pp. 2530-2540, Apr. 2005.
[9] S. Plyasunov and A. Arkin, "Efficient Stochastic Sensitivity Analysis of Discrete Event Systems," *J. Computational Physics,* vol. 221, pp. 724-738, 2007.
[10] M. Komorowski, J. Zurauskiene, and M. Stumpf, "Stochsens - Matlab Package for Sensitivity Analysis of Stochastic Chemical Systems," *Bioinformatics,* vol. 28, no. 5, pp. 731-733, 2012.
[11] Y.M. Marzouk, H.N. Najm, and L.A. Rahn, "Stochastic Spectral Methods for Efficient Bayesian Solution of Inverse Probelms," *J. Computational Physics,* vol. 224, no. 2, pp. 560-586, 2007.
[12] Y.M. Marzouk and H.N. Najm, "Dimensionality Reduction and Polynomial Chaos Acceleration of Bayesian Inference in Inverse Problems," *J. Computational Physics,* vol. 228, no. 6, pp. 1862-1902, 2009.
[13] G.M. Suel, R.P. Kulkarni, J. Dworkin, J. Garcia-Ojalvo, and M.B. Elowitz, "Tunability and Noise Dependence in Differentiation Dynamics," *Science,* vol. 315, no. 5819, pp. 1716-1719, 2007.
[14] G. Suel, J. Garcia-Ojalvo, L. Liberman, and M. Elowitz, "An Excitable Gene Regulatory Circuit Induces Transient Cellular Differentiation," *Nature,* vol. 440, no. 23, pp. 545-550, 2006.
[15] H. Maamar, A. Raj, and D. Dubnau, "Noise in Gene Expression Determines Cell Fate in Bacillus Subtilis," *Science,* vol. 317, no. 5837, pp. 526-529, July 2007.
[16] D. Schultz, E.B. Jacob, J. Onuchic, and P. Wolynes, "Molecular Level Stochastic Model for Competence Cycles in Bacillus Subtilis," *Proc. Nat'l Academy of Sciences USA,* vol. 104, no. 45, pp. 17582-17587, 2007.
[17] D. Kim, B. Debusschere, and H. Najm, "Spectral Methods for Parametric Sensitivity in Stochastic Dynamical Systems," *Biophysics J.,* vol. 92, no. 2, pp. 379-393, Jan. 2007.
[18] D. Sivia, *Data Analysis: A Bayesian Tutorial.* Oxford Science, 1996.
[19] K. Sanft, S. Wu, M. Roh, J. Fu, R.K. Lim, and L. Petzold, "Stochkit2: Software for Discrete Stochastic Simulation of Biochemical Systems with Events," *Bioinformatics,* vol. 27, no. 17, pp. 2457-2458, 2011.
[20] N. Wiener, "The Homogeneous Chaos," *Am. J. Math.,* vol. 60, pp. 897-936, 1938.
[21] R. Ghanem and P. Spanos, *Stochastic Finite Elements: A Spectral Approach.* Springer Verlag, 1991.
[22] M. Rosenblatt, "Remarks on a Multivariate Transformation," *Annals of Math. Statistics,* vol. 23, no. 3, pp. 470-472, 1952.
[23] K. Sargsyan, C. Safta, R. Berry, J. Ray, B. Debusschere, and H. Najm, "Efficient Uncertainty Quantification Methodologies for High-Dimensional Climate Land Models," Sandia Technical Report SAND2011-8757, Nov. 2011.
[24] D.A. Cox, J. Little, and D. O'Shea, *Ideals, Varieties and Algorithms: An Introduction to Computational Algebraic Geometry and Commutative Algebra.* Springer, 1997.
[25] S.A. Smolyak, "Quadrature and Interpolation Formulas for Tensor Products of Certain Classes of Functions," *Soviet Math. Doklady,* vol. 4, pp. 240-243, 1963.
[26] *Markov Chain Monte Carlo in Practice,* W.R. Gilks, S. Richardson, and D.J. Spiegelhalter, eds., pp. 59-74. Chapman and Hall, 1996.
[27] C. Andrieu, N. de Freitas, A. Doucet, and M.I. Jordan, "An Introduction to MCMC for Machine Learning," *Machine Learning,* vol. 50, pp. 5-43, 2003.
[28] O. Le Maître, H. Najm, R. Ghanem, and O. Knio, "Multi-Resolution Analysis of Wiener-Type Uncertainty Propagation Schemes," *J. Computational Physics,* vol. 197, pp. 502-531, 2004.
[29] K. Sargsyan, B. Debusschere, H. Najm, and O.L. Maître, "Spectral Representation and Reduced Order Modeling of the Dynamics of Stochastic Reaction Networks via Adaptive Data Partitioning," *SIAM J. Scientific Computing,* vol. 31, no. 6, pp. 4395-4421, 2010.
[30] X. Wan and G.E. Karniadakis, "An Adaptive Multi-Element Generalized Polynomial Chaos Method for Stochastic Differential Equations," *J. Computational Physics,* vol. 209, pp. 617-642, 2005.
[31] J. Gonzalez, I. Rojas, H. Pomares, J. Artega, and A. Prieto, "A New Clustering Technique for Function Approximation," *IEEE Trans. Neural Networks,* vol. 13, no. 1, pp. 132-142, Jan. 2002.
[32] F. Campolongo, A. Saltelli, T. Sørensen, and S. Tarantola, "Hitchhiker's Guide to Sensitivity Analysis," *Sensitivity Analysis,* A. Saltelli, K. Chan, and E. Scott, eds., Wiley, 2000.
[33] I.M. Sobol, "Sensitivity Estimates for Nonlinear Mathematical Models," *Math. Modeling and Computational Experiment,* vol. 1, pp. 407-414, 1993.

Khachik Sargsyan received the BS degree in applied mathematics and physics from Moscow Institute of Physics and Technology in 2002 and the PhD degree in applied and interdisciplinary mathematics from the University of Michigan, Ann Arbor, in 2007. He is a senior member of the technical staff at Sandia National Laboratories in Livermore, California. He has joined Sandia in 2007 and has since worked on a variety of projects. His research interests include statistical analysis, numerical methods for uncertainty quantification in physical systems with applications in biochemistry, statistical physics, population dynamics, climate science, and image recognition.

Cosmin Safta received the BS and MS degrees in aerospace engineering from the University "Politehnica" of Bucharest in 1995 and 1996, respectively, and the PhD degree in mechanical engineering from the University at Buffalo, SUNY, in 2004. He is a senior member of the technical staff at Sandia National Laboratories in Livermore, California. Before joining Sandia in 2007, he was with TTC Technologies in Long Island, New York, on the development of high-order methods for multidisciplinary CFD prediction tools. His academic background is in computational fluid dynamics and combustion, but his current research interests include uncertainty quantification, Bayesian analysis, climate modeling, and classification methods in epidemiology.

Bert Debusschere received the BS degree in mechanical engineering from the Katholieke Universiteit Leuven, Belgium, and the MS and PhD degrees in mechanical engineering from the University of Wisconsin, Madison, Wisconsin. He has been with Sandia National Laboratories since 2001. His main areas of research are uncertainty quantification, analysis of stochastic dynamical systems, Bayesian methods for parameter estimation, stochastic multiscale simulations, and model reduction for stiff chemical systems. In his work, he has studied a variety of systems ranging from atomistic flow simulations, and human immune system signaling pathways, to continuum microfluidics, combustion, and up the scale to climate models.

Habib Najm received the BE degree in mechanical engineering from the American University of Beirut in 1983, and the MS and PhD degrees in mechanical engineering from MIT in 1986 and 1989, respectively. He is a distinguished member of the technical staff at Sandia National Laboratories in Livermore, California. He was with the semiconductor process laboratory at Texas Instruments from 1989 through 1993, before joining the Combustion Research Facility at Sandia National Laboratories. His work at Sandia involves a range of computational science research, including development of numerical methods and computational tools for modeling and analysis of reacting flow, and the development of uncertainty quantification methods and tools, with application in general computational models. He is a coauthor of more than 70 archival journal articles and 11 US patents.