Spectral methods for prediction uncertainty quantification in Systems Biology

Anna Deneer, Jaap Molenaar, Christian Fleck

April 1, 2024

Abstract

Uncertainty is ubiquitous in biological systems. These uncertainties can be the result of lack of knowledge or due to a lack of appropriate data. Additionally, the natural variability of biological systems caused by intrinsic noise, e.g. in stochastic gene expression, leads to uncertainties. With the help of numerical simulations the impact of these uncertainties on the model predictions can be assessed, i.e. the impact of the propagation of uncertainty in model parameters on the model response can be quantified. Taking this into account is crucial when the models are used for experimental design, optimisation, or decision-making, as model uncertainty can have a significant effect on the accuracy of model predictions. We focus here on spectral methods to quantify prediction uncertainty based on a probabilistic framework. Such methods have a basis in, e.g., computational mathematics, engineering, physics, and fluid dynamics, and, to a lesser extent, systems biology. In this chapter, we highlight the advantages these methods can have for modelling purposes in systems biology and do so by providing a novel and intuitive scheme. By applying the scheme to an array of examples we show its power, especially in challenging situations where slow converge due to high-dimensionality, bifurcations, and spatial discontinuities play a role.

1 Introduction

Every mathematical model in systems biology is subject to uncertainty and incomplete knowledge [1–4]. This can be in the form of unknown model structure, unknown model parameters and imperfect experimental data. Characterizing and quantifying these sources is crucial, as the uncertainty can translate into inaccuracies in the model predictions. Information about the quality of model predictions is vital when applied as support for decision-making or optimization routines such as experimental design and parameter estimation [5]. The aim of uncertainty quantification (UQ) is to determine the likeliness of certain outcomes, given that some aspects of the system under study are not (exactly) known.

Generally, uncertainty is distinguished into two classes [6–8]. The first class is so-called aleatoric uncertainty. Aleatoric uncertainty stems from the intrinsic variability found in the system under consideration, for this reason it is also referred to as statistical uncertainty. For example, in the case of parameter estimation, this uncertainty is related to the fact that parameters may essentially vary over the system components (e.g., cells) [9], so that for the system as a whole only a distribution of parameter values can be estimated, and not one precise value per parameter. Gene expression noise is often a source of varying conditions (e.g., initial conditions or protein production rates) in cells resulting in variability of process parameters such as the protein production rate [9–13].

In contrast, the second class of uncertainty, termed epistemic (or systemic) uncertainty, is caused by a lack of information [6–8]. For example, in the case of parameter estimation, this may be caused by imperfect data sets that contain noisy, incoherent, or missing data points [14]. In such cases, the uncertainty can be reduced by performing extra experiments [15–17].

In biological systems both types of uncertainty are typically present [1]. In terms of modelling, both are usually dealt with by employing a probabilistic framework [18], in which model parameters can be represented according to a probability density function (PDF) [4, 6]. The choice of the type of PDF and the corresponding distribution of parameters is usually based on previous knowledge. For example, the case of a completely unknown parameter could be described by a uniform distribution on a broad (positive) interval. In other cases a parameter could be known to follow a normal or lognormal distribution with known mean and variance, established in previously performed experiments [19].

Among the field of UQ, Monte Carlo (MC) methods are most commonly used [20, 21]. In a MC approach the parameter PDFs are sampled and model responses for each sample recorded, thus providing a distribution of model outcomes and an indication of the uncertainty therein (e.g. by analyzing the distribution moments). These methods are simple in their implementation and are widely applicable. However, for models that have a large number of parameters or are computationally expensive, these MC procedures are often not feasible [6, 8].

As an alternative to MC, meta-modelling techniques are frequently adopted to deal with models that would otherwise be intractable. Support vector machines [22], artificial neural networks [23] and Bayesian networks [24] are examples of surrogate- or meta-modelling techniques used in systems biology. In this work, we focus on stochastic spectral methods, in particular polynomial chaos expansion (PCE), an approach that is widely used in engineering systems [7, 8] and to a lesser extent in biological systems [25–28] for UQ purposes. The aim of these approaches is to represent the model response as a series expansion. The advantage of this representation is that an approximation of the model response is obtained for *all* values of the uncertainty parameters. This form allows

immediate evaluation of statistics of the model outcome, either analytically or through sampling of the stochastic parameters, which can be done significantly faster than through MC methods for models that are problematic and computationally expensive [29].

This advantage comes at the cost of the need to calculate expansion coefficients. For this, two classes of spectral methods are in use. In the first class the governing equations of the model are reformulated such that each variable is represented by a spectral expansion. This results in a system of differential equations for the expansion coefficients and is known as intrusive spectral projection [6]. In the second class, consisting of so-called non-intrusive spectral projection approaches and followed in this paper, the expansion coefficients are determined using the model without changing the original model equations [30]. The advantage of this non-intrusive approach is that it requires only straightforward deterministic model evaluations and does not involve any reformulation of the model. It is particularly attractive in case of very large models where intrusive methods would become too laborious.

In the past, PCE has been shown to converge very slowly or not at all for models involving non-smooth functions [31, 32]. This is indeed a critical challenge for biological models, which often show complex, non-linear behaviour such as bifurcations and spatial discontinuities. In this work, we provide a scheme for non-intrusive spectral projection that may overcome these problems. It is easy to implement and we show its power through applying it to a number of biological models. In this scheme it is straightforward to use different basis functions for the expansion, such as Haar wavelets in case of models with bifurcations. We also present a simple segmentation method resulting in a piecewise continuous approximation of the model at hand which helps to deal with complex models avoiding the necessity of high order expansions. The examples treated in this paper each have a specific problem to overcome.

2 Methods

2.1 Spectral expansion

Let us consider a model $\mathbf{\Omega}$ that depends on a vector of stochastic input parameters $\boldsymbol{\theta}$. The model response Y can be any chosen quantity, e.g. the concentration of one of the model components or a function thereof. The uncertainty parameters θ_l are assumed to be independently distributed, each with PDF $P_l(\theta_l)$. So,

$$Y = \mathbf{\Omega}(\boldsymbol{\theta}), \quad \boldsymbol{\theta} \sim P(\boldsymbol{\theta}), \tag{1}$$

where $P(\boldsymbol{\theta})$ is the joint probability density function (PDF) for all U uncertainty parameters: $P(\boldsymbol{\theta}) = \prod_{l=1}^{U} P_l(\theta_l)$. The case of correlated parameters is discussed in Sec. 2.2.3. For reasons of clarity, we restrict here the explanation to Y and θ being scalar functions. In the section Practical Aspects we show how to deal

with more than one uncertainty parameter. Note that for correlated parameters the PDF would follow from the corresponding multivariate distribution.

The underlying model could be of any type, e.g., an ODE, a PDE, an algebraic, or a statistical model. This implies that Y may also depend on time and space. The challenge is to analyse the behaviour of Y as a function of $\boldsymbol{\theta}$. In cases where the numerical evaluation of the underlying model takes a considerable amount of computational time this tends to obstruct any form of comprehensive analysis. In this paper we present the use of a method that aims at making this tractable. The idea is to replace the original model by a meta-model, which is achieved by representing the output Y in terms of an expansion. This meta-model can be constructed such that it represents the underlying model to a high degree of accuracy, with the advantage of being much faster to evaluate than the original model.

This meta-model can be used to determine the distribution of the model response Y or reconstruct the function accurately at given points in the parameter space. In the spectral approach the model response is represented by

$$Y = \sum_{n=0}^{\infty} c_n \phi_n(\theta), \tag{2}$$

where c_n are the expansion coefficients (which can be time and/or space dependent) and ϕ_n are functions that are orthonormal with respect to the distribution of the uncertainty parameters as weight functions for an inner product, as dealt with in S1 Appendix A. For example, suitable basis functions for uniformly distributed parameters are Legendre polynomials, while for normally distributed parameters Hermite polynomials qualify. For any practical purpose the expansion needs to be truncated to a certain degree:

$$Y^s = \sum_{n=0}^{N} c_n \phi_n(\theta), \tag{3}$$

where N is the truncation degree. So, for such a meta-model Y^s N expansion coefficients have to be calculated. The advantage of (3) is that the statistics of the model response Y can be evaluated very fast, either analytically or through sampling of the parameters θ. The main computational cost of the expansion comes from the computation of the coefficients c_n. Below, we provide an easy-to-implement scheme for the calculation of these coefficients.

We use a non-intrusive approach to the spectral expansion, i.e., we treat the model equations as a black box, not requiring any tailoring to the equations to include the parameter uncertainties. The most commonly used method for non-intrusively determining the coefficients is through Gaussian quadrature schemes [6]. Here, we propose an alternative scheme. It is applicable to any set of orthonormal functions, allowing the flexibility to tackle different modelling

challenges. A key feature in the scheme is the introduction of the symmetric matrix:

$$\hat{B}_{n,m} = \int \phi_n(\theta)\, \theta\, \phi_m(\theta)\, P(\theta)\, d\theta. \tag{4}$$

Its eigenvalues $\lambda^{(l)}, l = 1, 2, \ldots,$ are real and its eigenvectors $u^{(l)}$ orthonormal. In the appendix we show that these eigenvalues and eigenvectors can be used to derive an expression for the coefficients c_n. After substitution, Eq. (3) then reads as

$$Y^s(\theta) = \sum_{l=1}^{N} Y(\lambda^{(l)})\, u_1^{(l)}\, \psi_l^s(\theta), \tag{5}$$

where

$$\psi_l^s(\theta) \equiv \sum_{n=0}^{N} u_{n+1}^{(l)}\, \phi_n(\theta). \tag{6}$$

The striking point here is that this expansion requires to evaluate the model only N times, namely for each of the eigenvalues $\lambda^{(l)}, l = 1, \ldots, N$. Note that all terms in the expansion that do not depend on the uncertainty parameter θ can be calculated in advance, so once and for all. This saves computation time for any future application. For models that take a long time to evaluate the use of Eq. (5) is a very fast alternative, compared to e.g. a Monte-Carlo approach. What's more, statistical moments like mean and variance follow directly from the expansion coefficients [6], thus requiring no further calculation. Note that expansion Eq. (5) is only exact in the limit $N \to \infty$. Taking a finite value for N introduces an inaccuracy. Therefore, N must be chosen with care and it is often not obvious beforehand which value of N will give reliable results. We will showcase in the examples underneath how one may deal with the choice of N. In the subsection Segmentation in the next section, we propose a very simple adjusted scheme to deal with cases where a high degree might result in infeasible computational times.

2.2 Practical aspects

Here, we treat some specific aspects of the method presented above.

2.2.1 PDFs and basis functions

We already mentioned Legendre and Hermite polynomials as typical basis functions for PCE. Legendre polynomials are defined over $[-1, 1]$ and are orthogonal with respect to the uniform distribution. The Hermite polynomials are defined over $\mathbb{R}$ and are orthogonal with respect to the Gaussian distribution. Both polynomials can be normalized with appropriate prefactors to ensure orthonormality. These two families of polynomials are most commonly used to represent biological parameters. The uniform distribution is typically applied in uninformed cases and the lognormal distribution in cases where there is prior information available on a parameter. In practice, sometimes other classical

orthogonal polynomial families are appropriate, e.g., Laguerre polynomials for Gamma distributions. Also, non-polynomial functions may be applied, such as spherical harmonics. The approach presented here can be used for any set of orthonormal functions.

The Legendre and Hermite polynomials are defined for standard uniform $\mathcal{U}(-1,1)$ and standard normal $\mathcal{N}(0,1)$ variables, respectively. In practice, the biological parameters are often not restricted to the corresponding intervals. In these cases we have to apply an isoprobabilistic transformation. For example, to obtain a normally distributed random variable k with mean μ and variance σ, so $k \sim \mathcal{N}(\mu,\sigma)$, from $\theta \sim \mathcal{N}(0,1)$, we need the transform

$$k = \mu + \sigma\theta, \qquad \theta \sim \mathcal{N}(0,1). \tag{7}$$

To obtain a uniformly distributed random variable k on the interval $[a,b]$, so $k \sim \mathcal{U}(a,b)$, from $\theta \sim \mathcal{U}(-1,1)$, we need the transform

$$k = (b+a)/2 + (b-a)\theta/2, \quad \theta \sim \mathcal{U}(-1,1). \tag{8}$$

A lognormally distributed random variable $k \sim \text{Lognormal}(\mu,\sigma)$ is obtained from $\theta \sim \mathcal{N}(0,1)$ via the transformation

$$k = \mu \exp\left[\alpha\theta - \frac{\alpha^2}{2}\right], \qquad \theta \sim \mathcal{N}(0,1), \tag{9}$$

where $\alpha = \sqrt{\ln\left(1 + \frac{\sigma^2}{\mu^2}\right)}$.

2.2.2 Multiple uncertainty parameters

Typically, biological models involve more than one random parameter, which means that the PCE basis $\{\phi_n(\boldsymbol{\theta}),\, n \in \mathbb{N}^M\}$ is multivariate. Extending (5) to the M-dimensional case is straightforward:

$$Y^s(\theta_1,\ldots,\theta_M) = \sum_{l_1=1}^{N} \cdots \sum_{l_M=1}^{N} Y(\lambda^{(l_1)},\ldots,\lambda^{(l_M)}) u_1^{(l_1)} \ldots u_1^{(l_M)} \psi_{l_1}^s(\theta_1) \ldots \psi_{l_M}^s(\theta_M). \tag{10}$$

Similar to MC approaches, PCE suffers from the curse of dimensionality [7, 33]. Note from (10) that the number of times the model has to be evaluated scales as N^M, where N is the expansion order and M the number of uncertainty parameters.

2.2.3 Correlated uncertainty parameters

To handle correlated parameters we consider the approach of transforming the basic functions. We discuss the case of two correlated parameters. In section S2 Supplementary Information we show the general case. Given that $P(\theta_1,\theta_2)$

is the distribution for the uncertainty parameters θ_1 and θ_2, $P(\theta_1)$ and $P(\theta_2)$ are the corresponding marginalised distributions, and $P(\theta_1, \theta_2) \neq 0, \forall \theta_1, \theta_2$, the new basis functions can be written as (for details see Supplementary Information S2):

$$\psi^s_{l_1,l_2}(\theta_1, \theta_2) = \sqrt{\frac{P(\theta_1)P(\theta_2)}{P(\theta_1, \theta_2)}} \psi^s_{l_1}(\theta_1)\psi^s_{l_2}(\theta_2), \tag{11}$$

where the functions ψ^s_n are orthonormal with respect to $P(\theta)$ and are given by Eq. 6. The expansion of a function depending on correlated parameters can be written as:

$$Y^s(\theta_1, \theta_2) = \sqrt{\frac{P(\theta_1)P(\theta_2)}{P(\theta_1, \theta_2)}} \sum_{l_1,l_2} Y(\lambda^{(l_1)}, \lambda^{(l_2)})\omega_{l_1,l_2}\psi^s_{l_1}(\theta_1)\psi^s_{l_2}(\theta_2), \tag{12}$$

with

$$\omega_{l_1,l_2} = \int \sqrt{P(\theta_1)P(\theta_2)P(\theta_1, \theta_2)}\psi^s_{l_1}(\theta_1)\psi^s_{l_2}(\theta_2)d\theta_1 d\theta_2. \tag{13}$$

2.2.4 Segmentation

In cases where an expansion to high order is needed to obtain the requested approximation accuracy or the model is computationally expensive, one needs to refer to optimised or effective methods [8, 14]. Models that show complex response surfaces (e.g., bifurcations) will require high order expansions to capture the inherent complexity. This is problematic since it not only requires to evaluate the model often, but also leads to time consuming summations in (10). To overcome these problems, we present here a simple and easy to implement scheme that segments the parameter intervals into subintervals, yielding a piecewise continuous approximation of the original function. Within each of these segments we then perform a separate expansion. In this approach we have to deal with a trade-off: the number of expansions is multiplied, but per expansion we have a (much) lower order of expansion. Below we argue why the second positive aspect greatly counterbalances the first negative aspect.
To determine the segments we define a scaling function g_m with $M \in \mathbb{N}$ and $L \in \mathbb{R}$ by:

$$\begin{aligned} g_m : [-L, L] \to \mathcal{I}_m &= \left[\frac{(2m-1)L}{2M+1}, \frac{(2m+1)L}{2M+1}\right], m \in \{-M, M\} \subset \mathbb{Z}, \\ \theta \mapsto g_m(\theta) &= \frac{2mL}{2M+1} + \frac{\theta}{2M+1}. \end{aligned} \tag{14}$$

This scaling function g_m divides the interval $[-L, L]$ into $2M + 1$ segments $\mathcal{I}_m$ of equal length. Of course, L must be larger than or equal to any value of θ. For example, consider the case $L = 1$. Then, the whole interval is $[-1, 1]$. For a segmentation granularity of $M = 1$, this interval is divided into three

subintervals: $\mathcal{I}_{-1} = [-1, -0.33]$, $\mathcal{I}_0 = [-0.33, 0.33]$ and $\mathcal{I}_{+1} = [0.33, 1]$. The expansion of Y on any subinterval $\mathcal{I}_m$ is given by:

$$Y(g_m(\theta)) = \sum_{l}^{N} Y(g_m(\lambda^{(l)}))u_1^{(l)}\psi_l(\theta). \tag{15}$$

Upon a variable transformation $y = g_m(\theta)$, Eq. 15 becomes:

$$Y(y) = \sum_{l}^{N} Y(g_m(\lambda^{(l)}))u_1^{(l)}\psi_l(g_m^{-1}(y)) : y \in \mathcal{I}_m. \tag{16}$$

After segmentation, the expansion on the interval $[-L, L]$ as a whole is a superposition of the expansions on the subintervals:

$$Y(y) = \sum_{m}^{M}\sum_{l}^{N} Y(g_m(\lambda^{(l)}))u_1^{(l)}\chi_m(y)\psi_l(g_m^{-1}(y)) : y \in [-L, L], \tag{17}$$

where $\chi_m(y)$ is an indicator function for selecting the correct segment:

$$\chi_m(y) = \begin{cases} 1 : y \in \mathcal{I}_m, \\ 0 : y \notin \mathcal{I}_m. \end{cases} \tag{18}$$

We can also define an index function to select $m^* \in \{-M, M\}$ for which $\chi_m(y) = 1$:

$$m^* = z(y) = \left\lfloor \frac{(2M+1)y}{2L} + \frac{1}{2} \right\rfloor \text{ for } |y| \leq L. \tag{19}$$

Using both index functions we can finally write the segmented reconstruction as:

$$Y(y) = \sum_{l}^{N} Y(g_{z(y)}(\lambda^{(l)}))u_1^{(l)}\psi_l(g_{z(y)}^{-1}(y)) : y \in [-L, L]. \tag{20}$$

Segmentation allows the use of lower order polynomials while maintaining the same accuracy (assuming a sensible choice for N and M) as a non-segmented higher order expansion. The number of model solutions required now scales as $(2M+1)^K N^K$, where K is the number of uncertainty parameters, M the segmentation granularity, and N the expansion order. The reduced accuracy by using a lower order expansion is compensated for by evaluating the model more often, as a result of zooming in. Expanding up to the N-th order for p uncertainty dimensions requires solving the system N^p times. Reconstruction requires the summation of N^{2p} terms. Therefore, it is advantageous to keep N as low as possible. Normally, the reconstruction error is large for low N, but this is mitigated by segmentation. Segmentation requires to evaluate the system $(2M+1)^p N^p$ times, but due to segmentation N can be taken much smaller.

To illustrate this with an example, we take a system with 2 species of interest and 5 uncertainty parameters θ_i. The expansion order is taken as $N = 8$. This

implies summing over $2 \times 8^{10} = 2,147,483,648$ terms per time point and per parameter set $(\theta_1, ..., \theta_5)$. In the case of segmented expansion, we can choose a lower N, for example $N = 3$ with a segmentation granularity of $M = 1$. The number of terms to be summed over is $2 \times 3^{10} = 118,098$. This is dramatically more efficient and stems from the fact that one only has to determine the segment in which the parameter set $(\theta_1, .., \theta_5)$ falls and choose the corresponding expansion coefficients. In the Results section we will present Example II in which segmentation indeed proves to be very beneficial.

2.2.5 Haar wavelet expansion

Traditional PCE methods are known to have difficulties with capturing discontinuous behaviour [31, 32]. Spectral convergence is only observed when solutions are sufficiently regular and continuous. Just like Fourier expansions, PCE suffers from Gibbs phenonema at discontinuities, resulting in slow convergence [8]. Haar wavelets have been suggested to overcome these difficulties [14, 32]. In contrast to global basis functions like the aforementioned polynomial systems, wavelet representations lead to localized decompositions, resulting in increased robustness at the cost of a slower convergence rate [8, 14]. Here, we discuss that Haar wavelets can be easily incorporated in the framework presented above and in Example IV underneath we show how they can be applied in practice.
As mother wavelet we take

$$\psi^W(y) = \begin{cases} 1 & \text{for } 0 \leq y < \frac{1}{2} \\ -1 & \text{for } \frac{1}{2} \leq y < 1 \\ 0 & \text{otherwise} \end{cases}. \tag{21}$$

By introducing a scaling factor j and a sliding factor k, we may construct the wavelet family

$$\psi^W_{j,k}(y) = 2^{j/2}\psi^W(2^j y - k), \qquad j = 0, 1 \ldots; \quad k = 0, \ldots, 2^j - 1. \tag{22}$$

Given the uncertainty parameter θ with its cumulative distribution function $F(\theta)$, we define the basis functions as

$$W_{j,k}(\theta) \sim \psi^W_{j,k}(F(\theta)). \tag{23}$$

By concatenating the indices j and k into one index $i \equiv 2^j + k$, we may expand the meta-model Y similarly as we did in Eq. (2):

$$Y(\theta) = \sum_{n=0}^{\infty} c_n \, W_n(\theta). \tag{24}$$

2.2.6 Sensitivity analysis

In sensitivity analysis one quantifies the effects of changes in the parameters on the variability of the model response. Here, we show how our PCE approach

allows for sensitivity analysis in an elegant way. In the case of local sensitivity analysis, small parameter variations around a certain point in parameter space are used to determine the effect on the model output [34]. This sensitivity is estimated via calculation of the partial derivatives of model output with respect to parameters, evaluated in that point [35]. Alternatively, global sensitivity approaches do not specify a specific point in parameter space [36]. For example, Sobol indices are a popular sensitivity measure as they provide a measure of global sensitivity and accurate information for most models [37]. Sobol indices are based on the decomposition of the variance of the output Y as a function of the contribution of the parameters (and possibly their combination), also called the ANalysis of VAriance, or ANOVA [37]. Thanks to the orthonormality of basis functions in PCE, Sobol indices can be determined analytically from the coefficients of the PCE [38, 39] So, once these coefficients are known, one gets the Sobol indices nearly for free. Given the PCE expansion of output Y, the total variance of the model output is given by

$$\hat{D} = \sum_{i \in I - \{0\}} c_i^2, \tag{25}$$

where I is the multi-index set of all variables and c_i the expansions coefficients. The 0th coefficient is not included as this is a constant. The partial variance is given by

$$D_{\theta_i} = \sum_{i \in I_{\theta_i}} c_i^2, \tag{26}$$

where I_{θ_i} is the multi-index set of parameter θ_i, i.e. where the ith term in the multi-index is larger than 0. The Sobol indices are then given by

$$S_{\theta_i} = \frac{D_{\theta_i}}{\hat{D}}, \tag{27}$$

In this way the relative contribution of parameter θ_i to the variance of the output is easily calculated.

2.3 Monte Carlo sampling

In the examples shown in this paper, we have compared PCE with Monte Carlo sampling. In these cases, we have used Quasi Monte Carlo methods using Sobol sequences, which show a faster rate of convergence than standard sequences of pseudorandom numbers [40]. In order to test for convergence, we have used the so-called blocking method. In the blocking method the error is estimated in a straightforward manner. The quantity of interest (i.e. the model response Y) is divided into several groups (or blocks). Then, for each block, we determine the moment of interest (e.g. mean or variance). The spread (variance) of values between blocks gives an estimate of the error. Finally, to test for convergence, we choose a threshold on the between-block variance. Note that the convergence rate of MC methods do not depend on the dimensionality of the parameter space that is being sampled, unlike PCE. MC merely scales by $\sqrt{N}$, where N is the number of samples.

2.4 Summary of implementation

In this section we provide an overview of the steps needed to arrive at a metamodel using PCE:

1. Determine which of the model parameters may show stochastic behaviour and decide upon an appropriate PDF for such parameters.

2. Choose a truncation degree N.

3. Based on the PDF in the previous steps, calculate the appropriate basis functions $\phi_n(\theta), n = 0, 1, 2 \ldots, N$.

4. Determine the $N \times N$ matrix $\hat{B}$ as defined in 4. For example, for Legendre polynomials $\hat{B}$ reads as

$$\hat{B}_{n,m} = \frac{n}{\sqrt{2n+1}\sqrt{2m+1}}\delta_{nm+1} + \frac{m}{\sqrt{2n+1}\sqrt{2m+1}}\delta_{nm-1}, \tag{28}$$

   where $n, m = 0, 1, 2 \ldots, N$.

   For Hermitian polynomials $\hat{B}$ reads as:

$$\hat{B}_{n,m} = \sqrt{n}\delta_{nm+1} + \sqrt{m}\delta_{nm-1}. \tag{29}$$

5. Calculate the eigenvalues $\lambda^{(l)}, l = 1, 2, \ldots, N$ and orthonormal eigenvectors $u^{(l)}$.

6. Calculate $\psi_l^s(\theta) = \sum_{n=0}^{N} u_{n+1}^{(l)}\, \phi_n(\theta)$.

7. Calculate $Y(\lambda^{(l)}), l = 1, 2, \ldots, N$ by evaluating the model N times.

8. Arrive at the metamodel $Y^s(\theta) = \sum_{l=1}^{N} Y(\lambda^{(l)})\, u_1^{(l)}\, \psi_l^s(\theta)$.

9. Eventually, apply post-processing through, e.g., sensitivity analysis.

3 Results

To test the performance of the present PCE approach in biological simulations, we have chosen four typical examples. Through these examples, we show how to deal with several challenges usually encountered in systems biology.

The first example has only one uncertainty parameter. Its simplicity allows comparison between the results of our approach with an exact solution.

The second example concerns a biochemical reaction network and is higher dimensional, i.e., it contains more than one uncertainty parameters. We use it to highlight the advantages of segmentation.

The third example is the glycolytic oscillator, which shows bifurcations, i.e., different dynamic behaviour for different parameter sets [41]. We use it to demonstrate the power of global sensitivity analysis, which in the PCE framework can be achieved without significant additional computational costs once the PCE coefficients have been calculated. In addition, this example allows us to show the use of mixed expansions, since the parameter PDFs follow different distributions. This leads to a combination of different families of basis functions, thus highlighting the flexibility of the PCE approach when applied to varying input uncertainties.

The last two examples have a spatial dimension. First, we consider the Schnakenberg model which is a well-known model of pattern formation and comes with challenges such as shifts from non-patterning to patterning regions [42]. In this example we demonstrate the advantage of using Haar wavelets over polynomial basis functions for systems with bifurcations. Second, we study a model describing pattern formation in plants, more specifically patterning of the hairs found on top of leaves, so-called trichomes [43]. In this model we show how to adapt the approach such that the computational costs are reduced as much as possible by carefully choosing the quantity of interest, without changing the standard set of steps.

3.1 Example I. Exponential decay: comparing performance of PCE to MC and an analytical solution

For this example case, we consider the extremely simple reaction system consisting of one decaying species:

$$A \xrightarrow{k} \emptyset.$$

Its dynamics are described by $A(t) = A_0 \exp^{-kt}$, where A_0 is the initial concentration of A at $t = 0$ and k the rate of decay. We test the PCE method against: *1)* the exact, analytical solution and *2)* the classical Monte Carlo approach.

The quantification of the sources of uncertainty constitute the second step in the analysis. This entails identification of the parameters that are unknown and modelling them in a probabilistic context. In this case, we assume that k is distributed according to a lognormal distribution with known mean and variance, i.e. $k \sim \text{Lognormal}(\mu, \sigma)$, and we choose $\mu = 0.5, \sigma = 0.2$. The PDF for k is shown in Figure 1 and the derivation for the exact PDF for A is given in the Supplementary Information.

Next, we determine how the uncertainty in k propagates through the model and affects concentration $A(t)$. To that end, we expand the function $A(t) = A_0 \exp^{-kt}$ in terms of Hermitian polynomials. Using (9), k is transformed into a standard normal variable θ. To arrive at the meta-model $Y^s(\theta)$ we truncate the expansion to a certain order N, as shown in (3). Choosing N is not straightforward and will involve some experimentation. In Figure 1 we compare

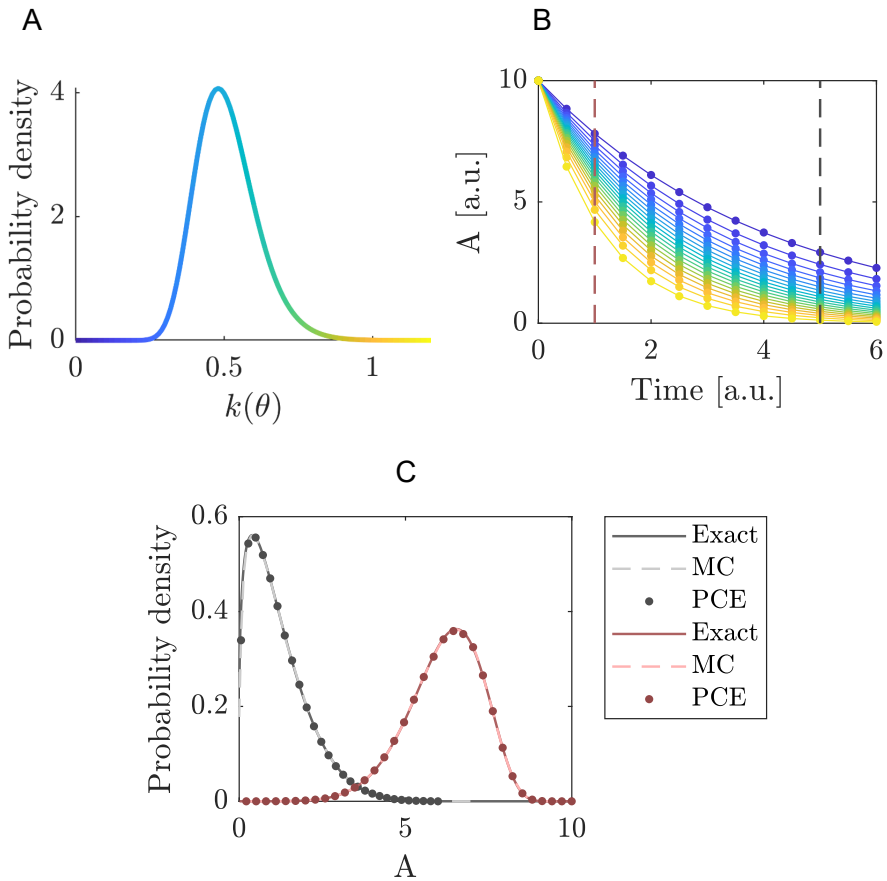

Figure 1: **Quantifying the uncertainty propagated by the decay rate in the exponential decay model.** A: Probability density function of the decay rate $k(\theta)$ with $\mu = 0.5$ and $\sigma = 0.2$. The colour gradient corresponds to the value of $k(\theta)$. B: The concentration of A up to $t = 6$ seconds. The solid lines indicate the analytical solution and the dots indicate the reconstruction using PCE with Hermite basis functions and an expansion order $N = 5$. The colour for each of the solutions correspond to the colour of the line in A, which indicates the value of $k(\theta)$ used for each of the depicted solutions. C: The dashed lines in B indicate a cross-section along the model response space at $t = 1$ for the red line, and at $t = 5$ for the grey line, determined through three methods. First, the exact dynamics of the model (solid lines), second through MC sampling using the exact model dynamics (dashed lines) and finally, through MC sampling of the reconstructed function as obtained through PCE (dots).

results for $N = 5$ to the analytical solution for different k values. This shows that for this expansion order the reconstructed function accurately matches the analytical solution.

In Figure 1, we focus on the distributions of $A(t = 1)$ and $A(t = 5)$. This is achieved by sampling the PCE using a large sample set χ of reduced (i.e. standard normally distributed) variables θ, $\chi_{sim} = \{\theta_j, j = 1, \ldots, n_{sim}\}$. The truncated series is then evaluated onto this sample: $Y^s_{sim} = \{\eta_j = \sum_{n=0}^{N} c_n \, \phi_n(\theta_j), j = 1, \ldots, n_{sim}\}$. These PDFs are obtained by kernel smoothing [44] using a sample set with $n_{sim} = 10^6$, drawn from the standard normal distribution with $\mu = 0$ and $\sigma = 1$. The kernel density estimator is given by

$$\hat{f}_Y(y) = \frac{1}{n_{sim} h} \sum_{j=1}^{n_{sim}} K\left(\frac{y - \eta_j}{h}\right), \tag{30}$$

with kernel function $K(t) = \frac{1}{\sqrt{2\pi}} \exp^{-t^2/2}$ and bandwidth h, which is determined by Silverman's rule of thumb [45]. Figure 1 shows that both MC and PCE perform well in reproducing the exact PDFs.

3.2 Example II. Biochemical reaction network: dealing with higher dimensions

In this example we present a simple model with multiple uncertainty parameters. The model describes the dynamics of two proteins x_1 and x_2 which bind together to form a dimer x_3. We consider the following reactions:

$$\emptyset \xrightarrow{k_1} x_1 \tag{31}$$

$$x_1 + x_2 \underset{k_3}{\overset{k_2}{\rightleftharpoons}} x_3 \tag{32}$$

$$\emptyset \xrightarrow{k_4} x_2 \tag{33}$$

$$x_1, x_2, x_3 \xrightarrow{k_5} \emptyset \tag{34}$$

In this network, the proteins x_1 and x_2 are produced at rates k_1 and k_4. Proteins x_1 and x_2 reversibly bind to form species x_3, with binding rate k_2 and unbinding rate k_3. All three proteins are degraded at the rate k_5.These interactions are visualized in a reaction scheme in Figure 2. The ODEs for this system are

$$\dot{x_1} = k_1 - k_2 x_1 x_2 + k_3 x_3 - k_5 x_1 \tag{35}$$

$$\dot{x_2} = k_4 - k_2 x_1 x_2 + k_3 x_3 - k_5 x_2 \tag{36}$$

$$\dot{x_3} = k_2 x_1 x_2 - k_3 x_3 - k_5 x_3 \tag{37}$$

We use for the parameters $k_1 - k_5$ log-normal distributions and expand the functions $x_1 - x_3$ in terms of Hermitian polynomials. In Figure 2 we compare the results of the segmented expansion with the non-segmented expansion and the exact results. For this comparison we chose the degree of expansion and

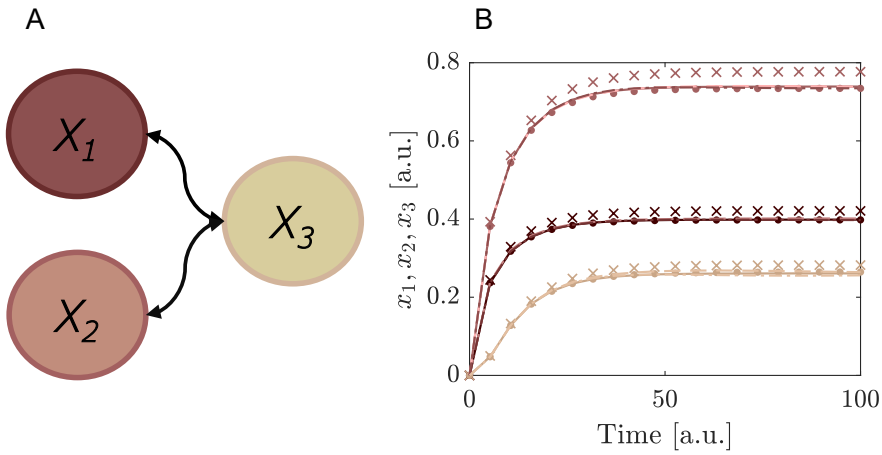

Figure 2: **Comparing the segmented expansion with the standard non-segmented expansion.** A: Interaction scheme of the model in Eqs.(35)-(37). Note that x_1, x_2 and x_3 all have the same degradation rate and x_1, x_2 have production rates in the model but this is not indicated in the scheme. B: Reconstruction of the system by a Hermitian expansion. For the segmented reconstruction we used $N = 2, M = 1$ (crosses) and $N = 3, M = 1$ (dots). For non-segmented expansion the expansion order was $N = 8$ (dashed lines) and $N = 9$, (dash-dotted lines). Note that these lines overlap with the true model solution (solid lines). We used two log-normal distributions with mean and standard deviation $\mu_1 = 0.1, \sigma_1 = 0.1$ for k_1, k_4, k_5 and $\mu_2 = 0.4, \sigma_2 = 0.1$ for k_2, k_3.

segmentation granularity such that the same number of model evaluations were required. We found that the subsequent summation to reconstruct the solutions for the differential equations improved by factors of 1000-30,000 when using the segmented expansion. See Table 1. As mentioned in the Methods section, this improvements stems from the large reduction in the number of terms to be summed over in the segmented case compared to the non-segmented expansion.

Dividing the parameter intervals into smaller sub-intervals is a relatively straightforward and simple way to circumvent huge computation times. Other, more intricate methods have been developed to tackle models with an even larger amount of parameters [33, 46, 47]. For example, using an adaptive algorithm that is based on classical statistical learning tools can result in a "sparse" PCE, that consists of only the significant coefficients in the expansion, thereby reducing the computational cost. This method has been tested on models of stochastic finite element analysis with up to 21 parameters [46].

Table 1: **Benchmarking of segmented and non-segmented expansion.** An overview of the number of model evaluations N_λ, the number of summation terms N_Σ and the time in seconds spent on summation (t_Σ), for different orders N of expansion, both segmented $M = 1$ and non-segmented $M = 0$. The last column highlights the speed-up factor when segmentation is used (keeping N_λ constant).

N	M	N_λ	N_Σ	$t_\Sigma[s]$	Fold-change
2	1	7776	3.07E+03	0.03	
3	1	59049	1.77E+05	0.17	
6	0	7776	1.81E+08	31.96	1065
9	0	59049	1.05E+10	5402	31776

3.3 Example III. Glycolytic oscillator: mixed input PDFs and global sensitivity

Living cells obtain energy by breaking down sugar in the biochemical process called glycolysis. In yeast cells, this glycolysis was observed to behave in an oscillatory fashion, where the concentration of various intermediates were increasing and decreasing within a period of several minutes [48]. This glycolytic oscillator can be modelled as a two-component system with a negative feedback [41]:

$$\dot{x} = -x + \alpha y + x^2 y \tag{38}$$

$$\dot{y} = \beta - \alpha y - x^2 y \tag{39}$$

where x and y are the concentrations of ADP (adenosine diphosphate) and F6P (fructose-6-phosphate) and α, β are kinetic parameters. Depending on the values of α and β the system will be in a stable limit cycle or a stable fixed point [41]. In this example, we assume α to be uniformly distributed on the interval $[0.1, 0.5]$ and β to follow a lognormal distribution with $\mu = 0.3$ and $\sigma = 0.1$. Because both uncertainty parameters come from a different distribution, the expansion will consist of multivariate polynomials $\Psi_{N,M}$ which are tensor products of the univariate polynomials. In this case, Legendre polynomials are used to expand $\alpha(\theta_1)$ and Hermite polynomials for $\beta(\theta_2)$, where $\theta_1 \sim \mathcal{U}(-1, 1)$ and $\theta_2 \sim \mathcal{N}(0, 1)$. This results in a mixed polynomial for the overall expansion as exemplified with a 3rd order Legendre polynomial $\tilde{L}_3$ and a 3rd order Hermite polynomial $\tilde{H}_3$, giving $\Psi_{3,3} = \tilde{L}_3 \cdot \tilde{H}_3$ (Figure 3).

The distributions of the uncertainty parameters were chosen such that they include the bifurcation point from stable limit cycle to the stable fixed point (Figure 3). For the purpose of this example we are interested in the concentration of y only and therefore reconstruct this model response using PCE. A good approximation is obtained with a truncation degree of the PCE of $N = 10$.

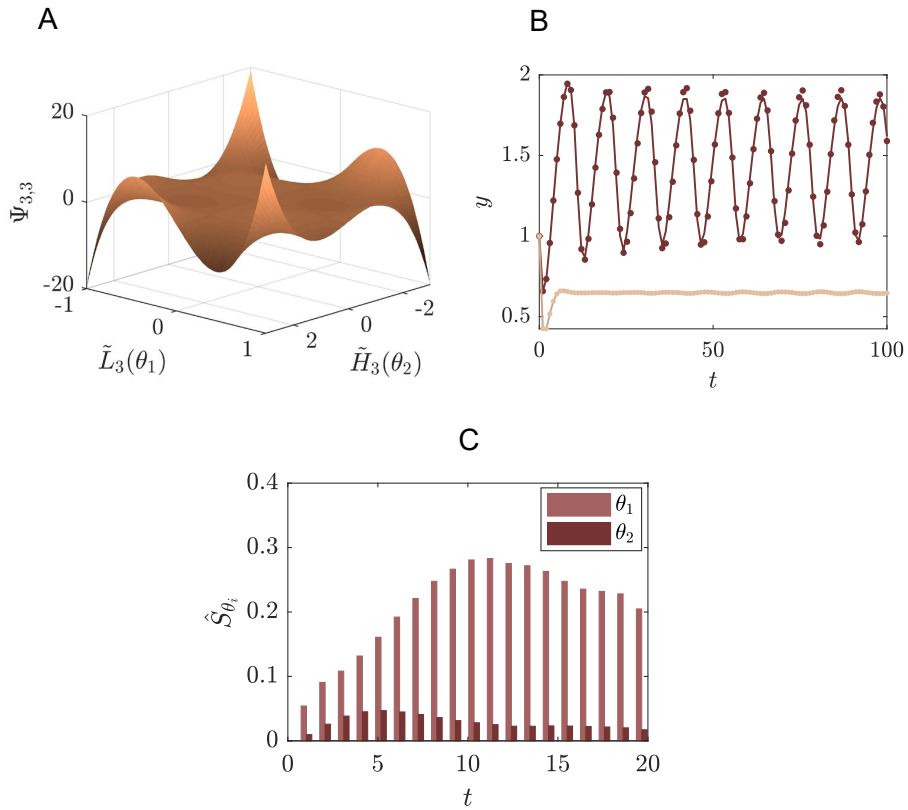

Figure 3: **Example of a system of glycolysis as defined in (38)-(39) where PCE is performed for the concentration of species y and for two uncertainty parameters.** A: Example of a multivariate polynomial, consisting of the tensor product between 3rd Legendre polynomial of the first random variable $\tilde{L}_3(\theta_1)$ and the 3rd Hermite polynomial of the second random variable $\tilde{H}_3(\theta_2)$. B: Solutions of concentration of fructose-6-phosphate (y) in the glycolytic oscillator model for two different points in the parameter space, obtained by solving the ODEs (lines) and reconstructing via PCE (dots). $\alpha = 0.1, \beta = 0.46$ produces an oscillation (dark coloured) whereas $\alpha = 0.5, \beta = 0.46$ gives a stable fixed point (light coloured). We used $N = 10$ as expansion order for the Legendre and Hermite polynomials. C: The first order Sobol sensitivity index $\hat{S}_{\theta_i}$ for the two random variables in the glycolytic oscillator model for the first 20 time points.

This value is relatively high, due to the bifurcation in the system. However, this case shows that convergence can be reached using PCE despite such challenges. Yet, the computational costs are still very tractable. In the following examples we deal with cases in which still higher order expansions are necessary due to non-smooth bifurcations.

In post-processing we may use the PC coefficients to determine the first order Sobol indices for the parameters α and β at each time point, providing a representation of the global sensitivity based on variance decomposition. The Sobol indices are readily available from the PC coefficients (see Methods). They have the advantage of being global measures of sensitivity. In Figure 3 we show the first order Sobol indices given in (26) for the first and second random variable. They indicate the contribution to the total output variance of either θ_1 or θ_2 individually. Higher-order terms would give an indication of interaction effects between θ_1 and θ_2, which are also readily available from the PCE coefficients but are not considered here for brevity.

3.4 Example IV. Schnakenberg model: dealing with spatial discontinuities

In this example we introduce a spatial component. We consider the Schnackenberg model, which is one of the simplest, but yet realistic two-species system that can produce periodic solutions and therefore has become a prototype for reaction diffusion systems. The Schnakenberg model consists of the following (dimensionless) equations [42]

$$\dot{u} = \nabla^2 u + \gamma(\alpha - u + u^2 v), \tag{40}$$

$$\dot{v} = d\nabla^2 v + \gamma(\beta - u^2 v). \tag{41}$$

α, β are reaction rates, γ a scale parameter and d the ratio of diffusion constants between the species u and v. The species u is auto-catalytically produced by the $u^2 v$ term in (40), whereby species v is consumed. There are certain combinations of the parameters α, β, γ and d for which the system will exhibit a stable pattern [49]; this region of parameter space is called the Turing space (TS). For the purpose of this example we limit the number of uncertainty parameters to one: the parameter α, fixing the other parameters at $\beta = 1, \gamma = 5$ and $d = 20$. We assume α to be distributed as $\alpha \sim \mathcal{U}(0.001, 0.45)$ and determine the TS for a range of α (Figure 4) using linear stability analysis (for details see [42, 49]). To that end, the model is simulated on a 1D grid of 20 cells. We focus on the concentration of species v at steady state and consider an expansion by both Legendre polynomials and Haar wavelets.

Polynomial chaos expansion is known for being inaccurate in regions that contain discontinuities [8, 31, 32]. In this example, the lack of convergence in PCE can be seen along the boundary of the patterning space (TS) in Figure 4, where the expansion by Legendre polynomials is indicated with the dashed lines.

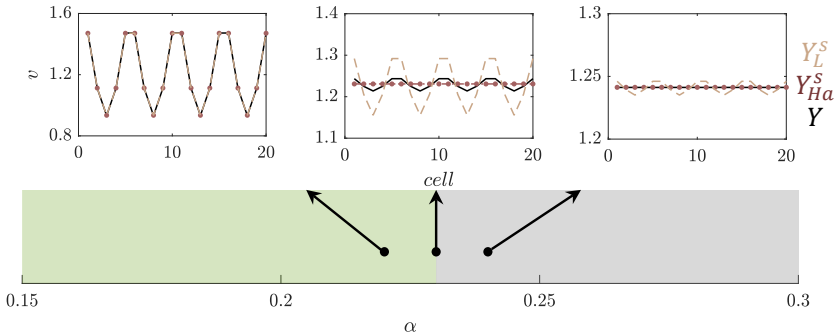

Figure 4: **Reconstruction of the concentration of species v of the Schnakenberg model, as defined in (41) at steady state, comparing Legendre and Haar wavelet expansions.** We consider the patterning of v in the Schnakenberg model for $\alpha \in [0.15, 0.3]$, as indicated on the main x-axis and assuming $\alpha \sim \mathcal{U}(0.001, 0.45)$, while fixing $\beta = 1, \gamma = 5$ and $d = 20$. The colour along this axis indicates the region in the 1D-parameter space whether a stable pattern will form (i.e., α is inside the Turing Space (TS), indicated by green colouring), or a homogeneous spatial distribution of v (grey colour). We highlight three examples of the patterns formed for different values of α, one inside the TS (left inset, $\alpha = 0.22$), one close to the boundary (middle inset, $\alpha = 0.23$) and one outside the TS (right inset, $\alpha = 0.24$). Within these examples we compare the true solution Y (solid black line) to the reconstructed function of v by PCE in terms of Legendre polynomials Y_L^s by a segmented expansion with polynomial order $N = 18$ and segmentation granularity $M = 3$ (dashed line), or Haar wavelets Y_{Ha}^s (dots), using the first 128 wavelets in the expansion, i.e. $N = 6$ resolution levels.

For the reconstruction of concentration of v in terms of Legendre polynomials, we used a segmented expansion with $N = 18$ and a segmentation granularity of $M = 3$, leading to a total of 126 model evaluations used in the expansion. To show that Haar wavelets perform much better in such a region, we additionally do an expansion in terms of Haar wavelets. As resolution level we take $N = 6$, which means a total of $N_w = 128$ wavelets are used in the expansion. In Figure 4 the performances of Legendre polynomials and Haar wavelets are compared in the vicinity of $\alpha = 0.23$ (middle inset, Figure 4), showing that the Haar-wavelets provide an improvement in accuracy at the bifurcation point, while using the same number of model evaluations (i.e. the same amount of information and computational cost) for the expansion.

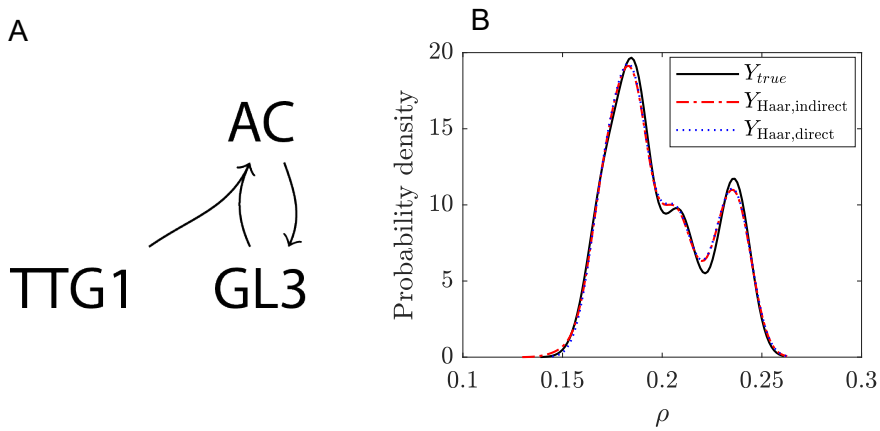

Figure 5: **Uncertainty quantification for the trichome system.** A: Schematic of the model. B: Probability density function of the trichome density in the Turing Space using either the indirect (dashed red line) or direct expansion (dotted blue line) method and for comparison the solution of the real model (solid black line). A resolution level of 3 (i.e., a total of 16 wavelets) has been used for the expansion.

3.5 Example V. Trichome patterning: dealing with spatial discontinuities

As an extra example of pattern formation we consider a model that describes trichomes. Trichomes are hairs found on the epidermal layer of leaves. In *Arabidopsis Thaliana* these trichomes form a regular pattern, where each trichome is separated by around three to four epidermal cells [50]. The model studied here consists of three proteins and their interactions which, taken together, can explain features of trichome patterning [43, 51]. Protein TRANSPARARANT TESTA GLABRA1 (TTG1) binds to the transcription factor GLABRA3 (GL3) which together form a trichome-promoting complex, called the activating complex (AC) [43]. Experimental data suggests that TTG1 is depleted from cells neighbouring a trichome [43]. For this reason the interaction between TTG1 and GL3 is modelled in a substrate-depletion form (Figure 5), where TTG1 acts as a substrate for the formation of AC [43]. After non-dimensionalisation this model consists of four parameters, none of which have been experimentally determined, highlighting the substantial amount of uncertainty within this model [52, 53]. Here, we examine the propagation of uncertainty in the parameters to the predicted pattern.

The trichome patterning is described by the following set of coupled ODEs

[43]:

$$\dot{\mathrm{TTG1}}_j = \alpha - \lambda \mathrm{TTG1}_j - \mathrm{TTG1}_j \mathrm{GL3}_j + \delta \hat{L} \mathrm{TTG1}_j \tag{42}$$

$$\dot{\mathrm{GL3}}_j = \beta \mathrm{AC}_j^2 - \mathrm{GL3}_j - \mathrm{TTG1}_j \mathrm{GL3}_j \tag{43}$$

$$\dot{\mathrm{AC}}_j = \mathrm{TTG1}_j \mathrm{GL3}_j - \mathrm{AC}_j \tag{44}$$

where α, λ, δ and β are parameters in the model and $\hat{L}$ describes the coupling between the cells. The subscript j indicates the jth cell. We solve these equations for 400 cells, grouped on a hexagonal grid of 20 by 20 cells.

In this example we focus on the parameter α, the basal production for TTG1. We assume this parameter to be uniformly distributed on the interval [0.4, 0.9]. We are interested in the number of trichomes that are predicted by the model, therefore we consider the trichome density ρ (total number of trichomes divided by the total number of cells in the simulated tissue) as the model response of interest. The number of trichomes is determined by simulating the system until steady state is reached and counting the number of cells for which the concentration of AC exceeds a threshold. The amount of AC is considered to be an indicator for trichome cell fate in plants, however the biological threshold for this is unknown. We set this threshold to the half-maximum of AC in the system. This leads to the following description of trichome density:

$$\rho = \frac{|\mathcal{T}|}{N} \tag{45}$$

$$\mathcal{T} = \left\{ j \in \mathcal{J} \;\middle|\; \mathrm{AC}_j \geq \frac{1}{2}\mathrm{AC}_{\max} \right\} \tag{46}$$

where $\mathcal{T}$ is the set of cells which exceed the AC threshold, $\mathcal{J}$ is the set of all cells on the grid, N is the total number of cells and $|\mathcal{T}|$ is the cardinality of $\mathcal{T}$.

The goal is to determine the uncertainty in ρ as a result of the uncertainty in α. To this end, we employ two different approaches. For both approaches we first transform α to a standard uniform variable by $\alpha = T^{-1}(\theta)$, using the transform function for a uniform variable given in (8). The first approach, refered to as the indirect approach, is the same as used in Example IV. To reconstruct the concentration at steady state for all cells, we expand the concentration of AC using Haar wavelets. From the result we may determine ρ. In this process we discriminate between cases where there is a pattern and where there is no pattern. For the latter, we need not solve the system as $\rho = 0$. Through linear stability analysis we determine beforehand whether a pattern will form or not, i.e., whether the chosen parameter set is in the Turing Space (TS) [42]. For a certain realisation θ we can determine ρ by

$$\rho(\theta) = \begin{cases} \frac{|\mathcal{T}^s|}{N} & \text{if } \theta \in \mathrm{TS} \\ 0 & \text{if } \theta \notin \mathrm{TS} \end{cases} \tag{47}$$

where $\mathcal{T}^s$ is the set of trichomes as determined from the reconstructed AC concentration profile.
Our second, direct approach is to directly reconstruct ρ as

$$\rho^s(\theta) = \sum_{l=1}^{N} Y(\lambda^{(l)})\, u_1^{(l)}\, \psi_l^s(\theta) \tag{48}$$

Similarly, as we did for $\rho(\theta)$ we can define $Y(\lambda^{(l)})$ as

$$Y(\lambda^{(l)}) = \begin{cases} \frac{|\mathcal{T}|}{N} & \text{if } \lambda^{(l)} \in \text{TS} \\ 0 & \text{if } \lambda^{(l)} \notin \text{TS} \end{cases} \tag{49}$$

In other words, we only solve the system and determine the trichome density if the parameter set falls within the Turing space. This lends robustness to the PCE for the non-smooth parts of the function $Y(\theta)$ and at the same time limits the amount of simulations to be performed, as the non-patterning parameter combinations need not be solved for.

We show here that there are multiple ways in which the uncertainty in the output can be captured. In this case of trichome patterning, we first tested an indirect method where the model output consists of concentration profiles from which the pattern features have to be extracted in post-processing, and secondly, we showed that the pattern feature could also be expanded directly, by taking the density as model output. Comparing the indirect and direct approaches, we conclude that both have similar levels of accuracy (Figure 5). Note that in both cases the expansions converge to the real solution at resolution level N = 6, which means a summation of 128 wavelets. The PDF in Figure 5 is constructed using 10^3 samples which costs 4.7 seconds for the wavelet reconstruction as opposed to 80.9 seconds for solving the full model a 1000 times in an MC approach.

3.6 Example VI. Plasmid transfection: dealing with correlated parameters

It can happen that in parameter space a structure occurs, i.e., that the multivariate joined probability cannot be written as a product of univariate distributions. In this section we exemplify how to handle such a case. As an example we choose the transfection of mammalian cells (e.g., Human embryonic kidney cells, HEK293) with two plasmids: plasmid pl_1 with the construct for induction (e.g. through chemical or optogenetics [54]), and plasmid pl_2 with the reporter construct. Denoting by n the number of pl_1 plasmids and by m the number of pl_2 plasmids in a specific cell, the corresponding reaction scheme reads:

$$\emptyset \xrightarrow{n\alpha I(t)} x_{n,m} \xrightarrow{\beta} \emptyset \tag{50}$$

$$\emptyset \xrightarrow{m\gamma x_{n,m}} y_{n,m} \xrightarrow{\mu} \emptyset \tag{51}$$

$$y_{n,m} \xrightarrow{\kappa} \qquad S \xrightarrow{\delta} \emptyset \tag{52}$$

where $x_{n,m}$ and $y_{n,m}$ are the concentrations of molecule x and y, resp., in a bacteria with plasmid composition nm. S is the concentration of the reporter molecule in the bulk. α and γ are the corresponding production rates per plasmid and β, μ are the degradation rates. $I(t)$ is the external induction signal. In what follows we will set $I(t) = 1$ for $t \geq 0$.

For sake of clarity we keep the system simple and ignore all complicating effects, such as gene expression noise, maturation of the reporter construct (e.g. GFP). However, for transient transfections of the mammalian cells, we need to take into account the distribution of plasmids among the cells, which is not constant due to cell division. The set of Ordinary Differential Equations describing the reactions given by 50- 52 read:

$$\dot{x}_{n,m} = n\alpha I(t) - \beta x_{n,m} \tag{53}$$

$$\dot{y}_{n,m} = \frac{m\gamma x_{n,m}}{1 + \eta x_{n,m}} - (\mu + \kappa) y_{n,m} \tag{54}$$

$$\dot{Q}_{n,m} = \omega \left(1 - \frac{\|Q\|_1}{N_{\max}}\right) \sum_{p,q} L^{p,q}_{n,m} Q_{p,q} \tag{55}$$

$$\dot{S} = v\kappa \sum_{n,m} Q_{n,m} y_{n,m} - \delta S. \tag{56}$$

The parameter η captures the saturation of the expression capacity of the cells. We set $\eta = 0.2\,V^{-1}$ throughout the calculations. V is the unit of the volume of the cells, e.g., $V = \mu m^3$. $Q_{n,m}$ is the number of cells that have n, m plasmids which are growing by ω. To follow the partition of plasmids upon division, the tensor L is introduced and described in further detail below. $\|Q\|_1 = \sum_{n,m} Q_{n,m} = N_{\text{tot}}$ which is the total number of cells at time t and $N_{\max}$ is the maximum number of bacteria due to environmental constraints. The growth rate is set to $\omega = 0.034\,h^{-1}$ based on a doubling time of 20 hours as being measured for HEK293T cells [55, 56]. S is the reporter (e.g. SEAP [57]) accumulated in the bulk due to secretion by the cells with rate κ. For non-secreted reporters (e.g. YFP) for single cell measurements, κ is set to zero. v is a volume correction factor, set to $v = 10^{-8}$.

We assume that the rate of dilution of the plasmids corresponds to the growth rate of the cells, i.e. plasmids are lost upon division. In reality, the plasmids are also degraded, but we assume the time scale of this degradation is much longer than the experimental duration and can therefore be ignored. To model the partition of plasmids upon cell division we assume that any partition of the number of plasmids is possible. The probability of finding a certain partition is given by the binomial distribution

$$P^k_n = \theta^k_n \binom{k}{n} (1-q)^{k-n} q^n \tag{57}$$

where n is the number of plasmids in the mother cell, k the number of plasmids in one of the daughter cells and q the probability, which for equal cell division is

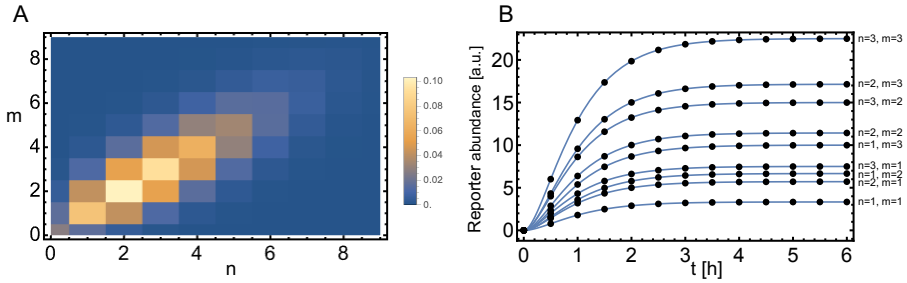

Figure 6: **Expansion with respect to a bivariate Poisson distribution.** A: Bivariate Poisson distribution given by Eq. 60 with $\lambda = 3$ and $\xi = 2.4$. B: Comparison of the exact solution for $y_{n,m}(t)$ given by Eq. 54 (solid lines) to the spectral expansion with $N = 6$ (dots). Parameters: $\alpha = 2\,(Vh)^{-1}$, $\beta = 2\,h^{-1}$, $\gamma = 6\,h^{-1}$, $\eta = 0.5\,V^{-1}$, $\mu = 1.5\,h^{-1}$, $\kappa = 0\,h^{-1}$.

$q = 0.5$. θ_n^k is the discrete unit step function with $\theta_n^k = 1$ for $k \geq n$ and $\theta_n^k = 0$ else. We can write the rate of change in the pools of cells as

$$\dot{Q}_{n,m} = 2\sum_{k,p} P_n^k P_m^p Q_{k,p} - Q_{n,m} = \sum_{k,p} \left(2P_n^k P_m^p - \delta_n^k \delta_m^p\right) Q_{k,p}. \tag{58}$$

$\delta_n^k = 1$ for $k = n$ and $\delta_n^k = 0$ else. From this, we can define the tensor used in Eq. 55:

$$L_{n,m}^{k,p} = \left(2P_n^k P_m^p - \delta_n^k \delta_m^p\right). \tag{59}$$

We treat the plasmid uptake as a Poisson process, i.e., the number of plasmids inside a bacteria is Poisson distributed. Assuming that the mean number of plasmid taken up is the same for pl_1 and pl_2 and a correlation exists between the uptake of two plasmids, the distribution is given by [58]:

$$P(n,m) = \frac{(\lambda-\xi)^m}{(\lambda-\xi)^n} \frac{(-1)^n \xi^n}{n!m!} U\left(-n, 1-n+m, -\frac{(\lambda-\xi)^2}{\xi}\right) e^{-2\lambda+\xi}, \tag{60}$$

where U is Kummer's confluent hypergeometric function, λ is the mean number of plasmids taken up, and $0 \leq \xi \leq \lambda$ is the correlation parameter. Note that for $\xi \to 0$ we find:

$$\lim_{\xi \to 0} P(n,m) = P(n)P(m), \quad \text{with: } P(x) = \frac{\lambda^x}{x!} e^{-\lambda}. \tag{61}$$

The correlation between n and m reads:

$$\text{corr}(n,m) = \frac{\text{cov}(n,m)}{\sqrt{\text{var}(n)\text{var}(m)}} = \frac{\xi}{\lambda}. \tag{62}$$

In Fig. 6A we show $P(n,m)$ for $\lambda = 3$ and $\xi = 2.4$. The asymmetry due to the correlation $\text{corr}(n,m) = 0.8$ can be clearly seen. The orthonormal basis function $\phi_p(n)$ with respect to the Poisson distribution read [59]:

$$\phi_p(n) = {}_2F_0\left(-p, -n; ; -\frac{1}{\lambda}\right)\sqrt{\frac{\lambda^p}{p!}}. \tag{63}$$

The matrix $\hat{B}$ defined in Eq. 4 can be calculated analytically and is given by:

$$\hat{B}_{n,m} = (n+\lambda)\delta_{nm} - \sqrt{n\lambda}\delta_{n-1m} - \sqrt{m\lambda}\delta_{nm-1}. \tag{64}$$

Note the dependence on the parameter λ.

We first consider a system in which the mammalian cells do not secrete the reporter molecule, e.g., YFP. In this case we set $\kappa = 0$, and consequently $S(t) = 0, \forall t$. In Fig. 6B we show $y_{n,m}(t)$ for different plasmid compositions and compare the expansion of $y_{n,m}(t)$ of the order $N = 6$ (dots) to the exact results (solid lines). A question of interest is what is the measured distribution of fluorescence intensities of the mammalian cells. The distribution of reporter molecule abundance is given by:

$$P(z,t) = \sum_{n,m}^{\infty} \delta_{z,\chi_{n,m}(t)} \frac{Q_{n,m}(t)}{\|Q\|_1}, \tag{65}$$

where $\chi_{n,m}(t)$ is a binning function: $\chi_{n,m}(t) = g\lfloor y_{n,m}(t)/g \rfloor$ (other binning methods are also possible, of course). Note that for sake of simplicity we ignore the transient phase after cell division in which the dynamic of $y_{n,m}$ adopts to the new, reduced plasmid composition. This is also a reasonable simplification considering that the steady state of $y_{n,m}$ is roughly reached after 4 h, as can be seen in Fig. 6B, in contrast to a cell doubling time of 20 h. In Fig. 7A we show the distribution $P(z, t = 50)$ for $g = 5$ for $\xi = 2.4$ (left) and $\xi = 0$ (right); we used the expansion with $N = 6$ for the calculations in Eq. 65.

Next, we consider a segregated reporter, e.g., SEAP. In this case the reporter abundance in the bulk is measured instead of single cell measurements. The temporal evolution of the bulk reporter is governed by Eq. 56. In Fig. 7B one can see that the dynamics of the reporter is clearly different due to correlation of the plasmid uptake.

3.7 Benchmarking of MC and PCE

To show the advantage of using PCE over MC, we turn to a trichome model that consists of 8 states and 31 parameters. We simulate this model on a grid of 35-by-35 hexagonal cells, resulting in a system of 9800 coupled ODEs [43]. We fix the initial conditions to the solution of the homogeneous steady state of a 'default' parameter set. We also use this parameter set to set the distributions for the uncertainty parameters, by choosing the mean to be the default

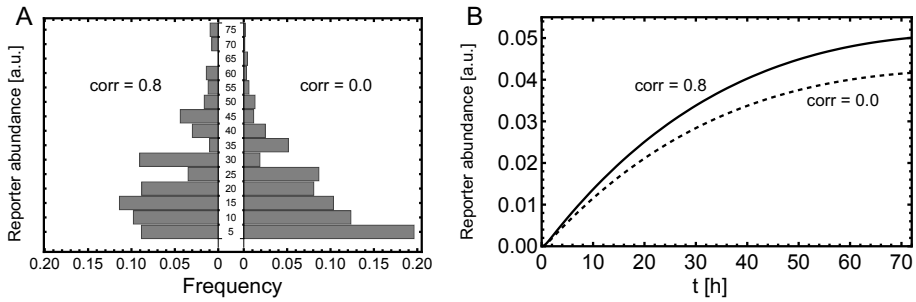

Figure 7: **Measurement of the reporter. Comparison of correlated and un-correlated plasmid uptake.** A: Histogram of the single cell readout, measured 6 h after induction. The parameters are the same as stated in Fig. 6 and the histogram is calculated according Eq. 65 with $g = 5$ B: Reporter abundance in the bulk measured for three days. The parameters are the same as in A, besides $\kappa = 3\,h^{-1}$, $\delta = 0.01\,h^{-1}$ and $v = 10^{-8}$. We used for the calculations the expansion of $y_{n,m}$ with $N = 6$.

value and the standard deviation to be 5% of the mean. In this benchmarking example, we aim to determine the mean and variance of the trichome density and do so by Monte Carlo simulation (as described in Methods) and PCE in terms of Hermite polynomials, with a varying number of uncertainty parameters. We start by assuming the first parameter θ_1 to following a lognormal distribution (see appendix for parameters and model equations), then in a second comparison we additionally assume θ_2 to follow a lognormal distribution (with different mean and variance), up to $\theta_1, \ldots, \theta_5$. For each of the different number of uncertainty parameters we determined the mean and variance, and compared the time computational time between MC and PCE (Figure 8). For both PCE and MC we used parallelisation, by running the model simulations on 9 cores in parallel (3.1 GHz CPU). We found convergence for MC at 1000 samples, which on average took between 75 to 80 minutes. For PCE we find a variety of computational times, ranging from around 2 minutes to 56 minutes for the lowest and highest number of uncertainty parameters, respectively. Note that for PCE a major computational advantage is the low cost of calculating the mean and variance. The mean follows directly from the zeroth coefficient: $\langle Y \rangle = c_0$ and the variance is given by:

$$\langle Y^2 \rangle = \sum_l \left| u_1^{(l)} \right|^2 Y^2(\lambda^{(l)}). \tag{66}$$

In comparing the mean and variance of the trichome density we find good agreement between MC and PCE (Table 2).

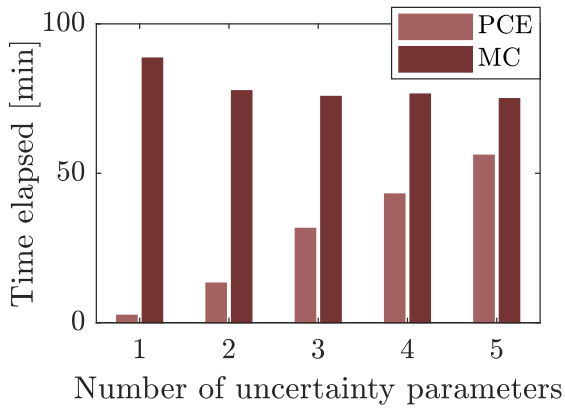

Figure 8: **Benchmarking of PCE and MC.** An overview of the computational times spent for PCE (light red) and MC (dark red).

Table 2: **Comparison of mean and variance found by MC and PCE.** NPar indicates the number of uncertainty parameters.

NPar	PCE order	$\langle Y \rangle$ (MC)	$\langle Y \rangle$ (PCE)	$\langle Y^2 \rangle$ (MC)	$\langle Y^2 \rangle$ (PCE)
1	30	0.033	0.033	3.80E-04	3.90E-04
2	10	0.034	0.031	6.15E-04	5.44E-04
3	7	0.034	0.030	7.26E-04	6.36E-04
4	5	0.034	0.030	7.19E-04	6.18E-04
5	4	0.034	0.029	7.22E-04	6.29E-04

4 Discussion and Conclusions

Through a series of examples we have presented an efficient and widely-applicable version of spectral methods for quantification of the effect of parameter uncertainty on model outcomes. The present scheme utilizes non-intrusive spectral projection based on polynomial functions or wavelets. The orthonormal properties of those functions provide a novel scheme to determine the expansion coefficients in a computationally fast way. The scheme is similar to the Golub-Welsh algorithm known from Gaussian quadrature [60]. In fact, the points indicating the roots of the polynomials used obtained in quadrature procedures, exactly correspond to the eigenvalues obtained from the matrix, which plays a central role in our approach. Quadrature methods as well as sampling methods are traditionally used to determine the coefficients of expansions of functions one is interested in [6]. The approach used here does effectively the same and provides an alternative to existing approaches, with the advantage that it is flexible and applicable to any set of orthonormal basis functions.

The method presented here requires no modification of the model equations. This is in general the main advantage of non-intrusive methods: there is no need to recast the model into a probabilistic framework. Instead, the random behaviour of parameters is accounted for through a set of deterministic simulations of the process for a restricted number of parameter values. These values are chosen such that they reflect the uncertainties in the parameters. To test the performance of our method we applied it to six different models: (1) a model of exponential decay, (2) a biochemical reaction network, (3) the glycolytic oscillator, (4) the Schnakenberg model, (5) a trichome model, and (6) transient mammalian plasmid uptake. Model (4) and (5) deal with spatial pattern formation. For each test case, the results of the PCE were compared to, if available, analytical solutions, non-PCE numerical simulations, or Monte Carlo simulations. In these comparisons we mostly focused on the accuracy of PCE. Even though the computational advantage is one of the major reasons for using PCE techniques, we do not focus on it because this aspect has already been extensively explored in various applications. For such comparisons we refer to other publications, see for example [6, 61, 62]. The logic behind the compututational advantages of PCE over MC extends to the methods we have presented here and is evident, e.g., in the case of trichome pattern formation (Example V), for which the output of the model was obtained 17 times faster when using PCE instead of MC.

The accuracy of the reconstruction by PCE depends on the choice of expansion order and basis functions. While the latter is determined by the PDFs of the input parameters, the choice of expansion order has to be chosen by the user. For example, in Examples I and III we chose $N = 5$ and $N = 10$, respectively. These choices were based on careful observation of the convergence properties of the method. In some cases the expansion order has to be chosen prohibitively large. For such a situation we propose an alternative approach that segments the parameter interval into subintervals, essentially zooming in on these subintervals such that a lower expansion order can be used in each sub-interval. In Example II we have shown that this segmentation approach can greatly reduce the computational costs in summation part of the expansion, thus providing a way to circumvent the curse of dimensionality. Such adaptations are required for the more difficult high-dimensional cases and the segmentation is a relatively simple and straight-forward method to tackle dimensionality problems yielding a piecewise continuous approximation to the original function. It is an alternative for so-called sparse PCE methods, that utilize only a small subset of the polynomial basis functions in order to limit the amount of model evaluations [33, 46, 47].

Convergence of the PCE may be poor in regions of the parameter space around a bifurcation [7]. In PCE, smooth polynomials are used in the expansions and they may show effects similar to the Gibbs phenomenon in Fourier expansions, i.e., the spectral basis is not suitable and leads to a slow convergence.

Since smooth functions like the Hermite and Legendre polynomials will fail to describe steep or discontinuous solutions, we explored the use of Haar wavelets. These wavelets lead to localized decompositions and this produces more robust behaviour [32]. We show in Example IV (Schnakenberg model) the advantages of using Haar wavelets over polynomials by focusing on the region in parameter space where the system transitions from spatially heterogeneous dynamics to spatially homogeneous dynamics take place. Around this bifurcation point an expansion in terms of Haar wavelets also shows slow convergence, but the accuracy of the expansion is 10 times better than when Legendre polynomials are used. For a fair comparison we kept the number of model evaluations in both approaches the same. In the vicinity of bifurcations Haar wavelets show greater robustness than the traditional polynomial basis functions. They thus provide a useful tool for biological systems which often feature such discontinuity.

In Example V (trichome pattern formation) we have highlighted the flexibility of the method: certain quantities, e.g., the scalar quantity of trichome density, can either be directly expanded or indirectly. By making use of that adaptability the number of model evaluations can be reduced while maintaining the same level of accuracy.

Finally, in Example VI we illustrate how to handle correlated parameters by means of correlated plasmid uptake by mammalian cells. We show how single-cell or bulk readout can be calculated using spectral expansion.

For a lower number of uncertainty parameters we find only small differences between the mean and variance determined by MC and PCE. As the dimensionality of the problem increases, PCE loses accuracy while showing increasingly higher computational times (Figure 2). Following the trend in Figure 2 for 6 uncertainty parameters, MC would likely be slightly faster. For such high-dimensional problems, the computational cost for PCE could be improved by e.g. sparse PCE [33, 46, 47], which has not been covered here for brevity but could be implemented within the presented scheme just as well. Furthermore, some accuracy could be gained by using Haar-wavelets instead of Hermite polynomials, which might improve the results around bifurcations in a patterning model such as the one presented here.

Overall, the approach presented here consists of a number of easy-to-implement steps and is widely applicable to a variety of systems which would be computationally costly when used in the context of uncertainty quantification. We believe that this approach provides a valuable tool in the tool-box for computational systems biology.

5 Funding

C.F. received funding from FET-Open research and innovation actions grant under the European Union's Horizon 2020 (CyGenTiG; grant agreement 801041).

6 CRediT authorship contribution statement

C.F. conceived the project, A.D. performed all calculations, A.D., J.M. and C.F. wrote the manuscript.

7 Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

8 Data availability

The code is publicly available at: https://github.com/AnnaDeneer/SpectralExpansion.

References

[1] Liesbet Geris, David Gomez-Cabrero, et al. *Uncertainty in biology.* Springer, 2016.

[2] 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 Comput Biol*, 3(10):e189, 2007.

[3] Eshan D Mitra and William S Hlavacek. Parameter estimation and uncertainty quantification for systems biology models. *Current opinion in systems biology*, 18:9–18, 2019.

[4] Simon van Mourik, Cajo Ter Braak, Hans Stigter, and Jaap Molenaar. Prediction uncertainty assessment of a systems biology model requires a sample of the full probability distribution of its parameters. *PeerJ*, 2:e433, 2014.

[5] Andreas Raue, Marcel Schilling, Julie Bachmann, Andrew Matteson, Max Schelke, Daniel Kaschek, Sabine Hug, Clemens Kreutz, Brian D Harms, Fabian J Theis, et al. Lessons learned from quantitative dynamical modeling in systems biology. *PloS one*, 8(9):e74335, 2013.

[6] Bruno Sudret. Uncertainty propagation and sensitivity analysis in mechanical models–contributions to structural reliability and stochastic spectral methods. *Habilitationa diriger des recherches, Université Blaise Pascal, Clermont-Ferrand, France*, 147, 2007.

[7] Roger Ghanem, David Higdon, and Houman Owhadi. *Handbook of uncertainty quantification*, volume 6. Springer, 2017.

[8] Olivier Le Maître and Omar M Knio. *Spectral methods for uncertainty quantification: with applications to computational fluid dynamics*. Springer Science & Business Media, 2010.

[9] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science (New York, NY)*, 297 (5584):1183–1186, August 2002. doi: 10.1126/science.1070919.

[10] Mukund Thattai and Alexander van Oudenaarden. Intrinsic noise in gene regulatory networks. *PNAS*, 98(15):8614–8619, July 2001. doi: 10.1073/pnas.151588598.

[11] W J Blake, M Kaern, C R Cantor, and J J Collins. Noise in eukaryotic gene expression. *Nature*, 422(6932):633–637, 2003. doi: 10.1038/nature01546.

[12] Lev S Tsimring. Noise in biology. *Reports on Progress in Physics*, 77(2): 026601–29, January 2014. doi: 10.1088/0034-4885/77/2/026601.

[13] Ilka Schultheiß Araújo, Jessica Magdalena Pietsch, Emma Mathilde Keizer, Bettina Greese, Rachappa Balkunde, Christian Fleck, and Martin Hülskamp. Stochastic gene expression in Arabidopsis thaliana. *Nature Communications*, 8(1):420–9, December 2017. doi: 10.1038/s41467-017-02285-7.

[14] T J Sullivan. *Introduction to Uncertainty Quantification*. Springer International Publishing, 2015. ISBN ISBN 978-3-319-23395-6.

[15] Julio R Banga and Eva Balsa-Canto. Parameter estimation and optimal experimental design. *Essays in biochemistry*, 45:195–209, 2008. doi: 10.1042/BSE0450195.

[16] Clemens Kreutz and Jens Timmer. Systems biology: Experimental design. *The FEBS journal*, 276(4):923–942, February 2009. doi: 10.1111/j.1742-4658.2008.06843.x.

[17] Tilman Barz, Harvey Arellano-Garcia, and Günter Wozny. Handling Uncertainty in Model-Based Optimal Experimental Design. *Industrial & Engineering Chemistry Research*, 49(12):5702–5713, June 2010. doi: 10.1021/ie901611b.

[18] Paul Kirk, Daniel Silk, and Michael PH Stumpf. Reverse engineering under uncertainty. In *Uncertainty in biology*, pages 15–32. Springer, 2016.

[19] Areti Tsigkinopoulou, Aliah Hawari, Megan Uttley, and Rainer Breitling. Defining informative priors for ensemble modeling in systems biology. *Nature protocols*, 13(11):2643–2663, 2018.

[20] Adrian Barbu and Song-Chun Zhu. *Monte Carlo Methods*. Springer Singapore, Singapore, 2020. ISBN 9789811329708 9789811329715. doi: 10.1007/978-981-13-2971-5.

[21] F James. Monte Carlo theory and practice. *Reports on Progress in Physics*, 43(9):1145–1189, 1980.

[22] William Stafford Noble et al. Support vector machine applications in computational biology. *Kernel methods in computational biology*, 71:92, 2004.

[23] Ryszard Tadeusiewicz. Neural networks as a tool for modeling of biological systems. *Bio-Algorithms and Med-Systems*, 11(3):135–144, 2015.

[24] Darren J Wilkinson. Bayesian methods in bioinformatics and computational systems biology. *Briefings in bioinformatics*, 8(2):109–116, 2007.

[25] Marc Martin-Casas and Ali Mesbah. Discrimination between competing model structures of biological systems in the presence of population heterogeneity. *IEEE life sciences letters*, 2(3):23–26, 2016.

[26] Joel A Paulson, Marc Martin-Casas, and Ali Mesbah. Fast uncertainty quantification for dynamic flux balance analysis using non-smooth polynomial chaos expansions. *PLoS computational biology*, 15(8):e1007308, 2019.

[27] Stefan Streif, Felix Petzke, Ali Mesbah, Rolf Findeisen, and Richard D Braatz. Optimal experimental design for probabilistic model discrimination using polynomial chaos. *IFAC Proceedings Volumes*, 47(3):4103–4109, 2014.

[28] Marissa Renardy, Tau-Mu Yi, Dongbin Xiu, and Ching-Shan Chou. Parameter uncertainty quantification using surrogate models applied to a spatial model of yeast mating polarization. *PLoS computational biology*, 14(5): e1006181, 2018.

[29] Dongbin Xiu and George Em Karniadakis. The wiener–askey polynomial chaos for stochastic differential equations. *SIAM journal on scientific computing*, 24(2):619–644, 2002.

[30] Michael Eldred, Clayton Webster, and Paul Constantine. Evaluation of non-intrusive approaches for wiener-askey generalized polynomial chaos. In *49th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, 16th AIAA/ASME/AHS Adaptive Structures Conference, 10th AIAA Non-Deterministic Approaches Conference, 9th AIAA Gossamer Spacecraft Forum, 4th AIAA Multidisciplinary Design Optimization Specialists Conference*, page 1892, 2008.

[31] Alexandre Joel Chorin. Gaussian fields and random flow. *Journal of Fluid Mechanics*, 63(1):21–32, 1974.

[32] OP Le Maıtre, OM Knio, HN Najm, and RG Ghanem. Uncertainty propagation using wiener–haar expansions. *Journal of computational Physics*, 197(1):28–57, 2004.

[33] Dongbin Xiu. Efficient collocational approach for parametric uncertainty analysis. *Communications in computational physics*, 2(2):293–309, 2007.

[34] Brian P Ingalls. *Mathematical Modeling in Systems Biology.* An Introduction. MIT Press, July 2013. ISBN 0-262-01888-8.

[35] Brian Ingalls. Sensitivity analysis: From model parameters to system behaviour. *Essays in Biochemistry*, 45:177–194, September 2008. ISSN 0071-1365, 1744-1358. doi: 10.1042/bse0450177.

[36] A. Saltelli, editor. *Global Sensitivity Analysis: The Primer.* John Wiley, Chichester, England ; Hoboken, NJ, 2008. ISBN 978-0-470-05997-5. OCLC: ocn180852094.

[37] Ilya M Sobol. Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Mathematics and computers in simulation*, 55(1-3):271–280, 2001.

[38] Bruno Sudret. Global sensitivity analysis using polynomial chaos expansions. *Reliability engineering & system safety*, 93(7):964–979, 2008.

[39] Géraud Blatman and Bruno Sudret. Efficient computation of global sensitivity indices using sparse polynomial chaos expansions. *Reliability Engineering & System Safety*, 95(11):1216–1229, 2010.

[40] IM Soboĺ. Quasi-monte carlo methods. *Progress in Nuclear Energy*, 24 (1-3):55–61, 1990.

[41] Steven H. Strogatz. *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering.* Studies in Nonlinearity. Addison-Wesley Pub, 1994. ISBN 978-0-201-54344-5.

[42] James D Murray. *Mathematical biology: I. An introduction*, volume 17. Springer Science & Business Media, 2007.

[43] Daniel Bouyer, Florian Geier, Friedrich Kragler, Arp Schnittger, Martina Pesch, Katja Wester, Rachappa Balkunde, Jens Timmer, Christian Fleck, and Martin Hülskamp. Two-dimensional patterning by a trapping/depletion mechanism: The role of TTG1 and GL3 in Arabidopsis trichome formation. *PLoS biology*, 6(6):e141, June 2008. doi: 10.1371/journal.pbio.0060141. ZSCC: NoCitationData[s0].

[44] Adrian W Bowman and Adelchi Azzalini. *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*, volume 18. OUP Oxford, 1997.

[45] Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.

[46] Géraud Blatman and Bruno Sudret. An adaptive algorithm to build up sparse polynomial chaos expansions for stochastic finite element analysis. *Probabilistic Engineering Mechanics*, 25(2):183–197, 2010.

[47] Fabio Nobile, Raúl Tempone, and Clayton G Webster. A sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM Journal on Numerical Analysis*, 46(5):2309–2345, 2008.

[48] Benno Hess and Arnold Boiteux. Mechanism of glycolytic oscillation in yeast, i. aerobic and anaerobic growth conditions for obtaining glycolytic oscillation. *Biological Chemistry*, 349(2):1567–1574, 1968.

[49] Ping Liu, Junping Shi, Yuwen Wang, and Xiuhong Feng. Bifurcation analysis of reaction-diffusion schnakenberg model. *Journal of Mathematical Chemistry*, 51(8):2001–2019, 2013.

[50] Martin Hülskamp. Plant trichomes: A model for cell differentiation. *Nature Reviews Molecular Cell Biology*, 5(6):471–480, June 2004. doi: 10.1038/nrm1404.

[51] Martina Pesch and Martin Hülskamp. Creating a two-dimensional pattern de novo during arabidopsis trichome and root hair initiation. *Current opinion in genetics & development*, 14(4):422–427, 2004.

[52] Marina Pesch and Martin Hülskamp. One, two, three... models for trichome patterning in arabidopsis? *Current opinion in plant biology*, 12(5): 587–592, 2009.

[53] Ben Scheres. Non-linear signaling for pattern formation? *Current opinion in plant biology*, 3(5):412–417, 2000.

[54] Naoto Yoshida and Misa Sato. Plasmid uptake by bacteria: A comparison of methods and efficiencies. *Applied Microbiology and Biotechnology*, 83(5):791–798, July 2009. ISSN 0175-7598, 1432-0614. doi: 10.1007/s00253-009-2042-4.

[55] Evan Tan, Cara Sze Hui Chin, Zhi Feng Sherman Lim, and Say Kong Ng. HEK293 Cell Line as a Platform to Produce Recombinant Proteins and Viral Vectors. *Frontiers in Bioengineering and Biotechnology*, 9:796991, December 2021. ISSN 2296-4185. doi: 10.3389/fbioe.2021.796991.

[56] Liao Zhang, Jianhui Gao, Xi Zhang, Xiaoyin Wang, Tianyun Wang, and Junhe Zhang. Current strategies for the development of high-yield HEK293 cell lines. *Biochemical Engineering Journal*, 205:109279, May 2024. ISSN 1369703X. doi: 10.1016/j.bej.2024.109279.

[57] Joel Berger, Joachim Hauber, Regina Hauber, Reinhard Geiger, and Bryan R Cullen. Secreted placental alkaline phosphatase: a powerful new quantitative indicator of gene expression in eukaryotic cells. *Gene*, 66(1): 1–10, 1988.

[58] Peter Berkhout and Erik Plug. A bivariate Poisson count data model using conditional probabilities. *Statistica Neerlandica*, 58(3):349–364, 2004. ISSN 1467-9574. doi: 10.1111/j.1467-9574.2004.00126.x.

[59] H. Ogura. Orthogonal functionals of the Poisson process. *IEEE Transactions on Information Theory*, 18(4):473–481, July 1972. ISSN 0018-9448. doi: 10.1109/TIT.1972.1054856.

[60] Gene H Golub and John H Welsch. Calculation of gauss quadrature rules. *Mathematics of computation*, 23(106):221–230, 1969.

[61] Samih Zein, Benoît Colson, and François Glineur. An efficient sampling method for regression-based polynomial chaos expansion. *Communications in computational physics*, 13(4):1173–1188, 2013.

[62] Noura Fajraoui, S Marelli, and B Sudret. On optimal experimental designs for sparse polynomial chaos expansions. *arXiv preprint arXiv:1703.05312*, 2017.