PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Fast uncertainty quantification for dynamic flux balance analysis using non-smooth polynomial chaos expansions

Joel A. Paulson, Marc Martin-Casas, Ali Mesbah*

Department of Chemical and Biomolecular Engineering, University of California, Berkeley, Berkeley, California, United States of America

* mesbah@berkeley.edu

OPEN ACCESS

Citation: Paulson JA, Martin-Casas M, Mesbah A (2019) Fast uncertainty quantification for dynamic flux balance analysis using non-smooth polynomial chaos expansions. PLoS Comput Biol 15(8): e1007308. https://doi.org/10.1371/journal.pcbi.1007308

Editor: Pedro Mendes, University of Connecticut School of Medicine, UNITED STATES

Received: February 14, 2019

Accepted: July 31, 2019

Published: August 30, 2019

Data Availability Statement: All relevant data are within the manuscript and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing interests: The authors have declared that no competing interests exist.

Abstract

We present a novel surrogate modeling method that can be used to accelerate the solution of uncertainty quantification (UQ) problems arising in nonlinear and non-smooth models of biological systems. In particular, we focus on dynamic flux balance analysis (DFBA) models that couple intracellular fluxes, found from the solution of a constrained metabolic network model of the cellular metabolism, to the time-varying nature of the extracellular substrate and product concentrations. DFBA models are generally computationally expensive and present unique challenges to UQ, as they entail dynamic simulations with discrete events that correspond to switches in the active set of the solution of the constrained intracellular model. The proposed non-smooth polynomial chaos expansion (nsPCE) method is an extension of traditional PCE that can effectively capture singularities in the DFBA model response due to the occurrence of these discrete events. The key idea in nsPCE is to use a model of the singularity time to partition the parameter space into two elements on which the model response behaves smoothly. Separate PCE models are then fit in both elements using a basis-adaptive sparse regression approach that is known to scale well with respect to the number of uncertain parameters. We demonstrate the effectiveness of nsPCE on a DFBA model of an *E. coli* monoculture that consists of 1075 reactions and 761 metabolites. We first illustrate how traditional PCE is unable to handle problems of this level of complexity. We demonstrate that over 800-fold savings in computational cost of uncertainty propagation and Bayesian estimation of parameters in the substrate uptake kinetics can be achieved by using the nsPCE surrogates in place of the full DFBA model simulations. We then investigate the scalability of the nsPCE method by utilizing it for global sensitivity analysis and maximum a posteriori estimation in a synthetic metabolic network problem with a larger number of parameters related to both intracellular and extracellular quantities.

Author summary

Construction and validation of mathematical models in biological systems involving genome-scale biomolecular networks is a challenging problem. This article presents a

novel surrogate modeling method that can accelerate parameter inference from experimental data and the quantification of uncertainty in the predictions of complex dynamic biological models, with a particular emphasis on nonlinear models with non-smooth behavior. The method is applied to infer extracellular kinetic parameters in a batch fermentation reactor consisting of diauxic growth of *E. coli* on a glucose/xylose mixed media as well as a larger synthetic metabolic network problem. The proposed approach enables rigorous quantification of parameter uncertainty to determine whether or not available data is sufficient for estimation of all unknown model parameters.

Introduction

The utility of mathematical modeling in biology is on the rise due to computational advancements and the increasing availability of data provided by high-throughput experimental techniques [1]. Flux balance analysis (FBA) is widely used for modeling cellular metabolism in a large range of metabolic and biochemical engineering problems [2, 3]. Given a constrained metabolic network, FBA assumes the intracellular fluxes are regulated by the cell to optimize a predefined cellular objective function (e.g., maximizing the biomass growth rate [4]) subject to mass balances of the intracellular metabolites and other feasibility constraints (e.g., bounds on the substrate uptake and product secretion rates). However, FBA only identifies metabolic flux distributions at steady-state and, thus, provides no information on metabolite concentrations or the dynamic behavior of the fluxes. A dynamic extension to FBA, commonly referred to as dynamic FBA (DFBA), was originally developed in [5] and has been subsequently applied in several applications [6–9]. In DFBA models, the intracellular fluxes are given by the solution of a FBA model, which is coupled to a set of dynamic equations that describes the time-varying nature of the extracellular substrate and product concentrations as a function of the extracellular environment [10]. The key assumption in DFBA is that the intracellular fluxes equilibrate instantaneously. This "quasi steady-state" assumption is valid as long as the intracellular dynamics are significantly faster than the extracellular dynamics.

Generally, the prediction of the behavior of biological systems such as those described by DFBA models can be subject to various sources of uncertainty including unknown model parameters, unknown model structure, and experimental uncertainty such as measurement error [11]. Accurate quantification of these uncertainties, as well as their impact on the quality of model predictions, is vital when applying these models in decision-support or optimization tasks such as parameter estimation or optimal experiment design. The task of uncertainty quantification (UQ) can be divided into two major problems: *forward* uncertainty propagation and *inverse* uncertainty estimation. The forward problem focuses on propagating all uncertainties through the model to predict the overall uncertainty in the outputs, whereas the inverse problem aims to calibrate the model with experimental data [12–14]. However, the most commonly used UQ methods are intractable for expensive-to-evaluate computational models [15], which has severely limited their application to DFBA models. An overview of the various challenges in DFBA simulations can be found in [16].

Surrogate modeling techniques are being increasingly adopted to enable complex UQ analyses that would otherwise be impossible. Of the available surrogate modeling approaches, polynomial chaos expansions (PCEs) are one of the most commonly used methods for UQ, which have been shown to yield accurate representations of model outputs using limited computational resources in various engineering systems [17–20] as well as biological systems [21–23]. However, an important underlying assumption in PCE is that the model response is a smooth

function of the uncertain parameters such that the response can be accurately approximated by a collection of polynomial functions. For non-smooth models, PCE has been shown to either converge very slowly or even fail to converge altogether depending on the type of non-smoothness [24, 25]. This is a critical challenge in DFBA models because they are known to become singular (i.e., lose differentiability) at certain time points due to the underlying quasi steady-state assumption [10, 26, 27], meaning that even state-of-the-art PCE methods are not directly applicable to DFBA models.

In this work, we propose an extension to PCE, referred to as non-smooth PCE (nsPCE), that can adequately capture the non-smooth behavior exhibited by DFBA models. The underlying concept behind the proposed nsPCE framework is that the time of occurrence of any singularity in a DFBA model is a smooth function of the parameters, which can be effectively modeled with a PCE. Thus, for any given time of interest, the PCE model of the singularity time can be used to partition the parameter space into two non-overlapping regions (or elements) that represent the collection of parameters for which the singularity has and has not occurred. Separate PCEs can then be constructed over each of these elements, leading to a piecewise polynomial approximation of the overall model response. We adopt a non-intrusive, regression-based approach for PCE construction from a limited number of expensive DFBA simulations. In particular, we take advantage of state-of-the-art sparse regression methods to systematically locate the terms that have the greatest impact on the model response out of a very large candidate set of terms. By exploiting sparsity, we can mitigate the curse-of-dimensionality that can plague traditional PCE, allowing the application of the proposed nsPCE approach to problems with reasonably large number of uncertain parameters.

To demonstrate the effectiveness of the nsPCE method, it is applied to accelerate Bayesian estimation of parameters in the substrate uptake kinetic expressions of diauxic growth of a batch monoculture of *Escherichia coli* on a glucose and xylose mixed media. The metabolic network reconstruction used for *E. coli* is iJ904, which is a genome-scale model that contains 1075 reactions and 761 metabolites [28]. Parameter estimation is performed using measurements of the concentrations of extracellular metabolites and biomass that are taken at certain time points throughout the batch. We selected this particular system due to the fact that reported parameter estimates were determined from experimental data using a trial-and-error procedure [8]. This was likely due to the computational complexity of the genome-scale DFBA model in conjunction with the limited data set that may not enable unique estimation of parameters. In addition, we demonstrate how nsPCE can be applied to vastly speedup forward UQ analyses including global sensitivity analysis and estimation of the probability distribution of the model response. To demonstrate the scalability of nsPCE, it is used for maximum a posteriori parameter estimation in a synthetic metabolic network problem with twenty unknown parameters related to quantities in both the intracellular reaction network and the extracellular environment. The codes that implement the proposed nsPCE method for generic DFBA models are provided at the repository [29].

Methods

Dynamic flux balance analysis models

We focus on modeling a microbial cultivation process using dynamic flux balance analysis (DFBA), in which the bioreactor is viewed as a combination of the fluid medium (extracellular environment) and the microorganisms (intracellular environment). Cell walls act as physical boundaries between these two phases, through which certain chemical metabolites are

exchanged. The DFBA model can be mathematically formulated as [26]

$$\dot{\mathbf{s}}(t) = \mathbf{f}(t, \mathbf{s}(t), \mathbf{v}(\mathbf{s}(t))), \quad \mathbf{s}(t_0) = \mathbf{s}_0, \tag{1}$$

with $\mathbf{v}(\mathbf{s}(t))$ being an element of the solution set of the flux balance model

$$\begin{aligned} \mathbf{v}(\mathbf{s}) \in \underset{\mathbf{v}}{\text{argmax}} \quad & h(\mathbf{v}, \mathbf{s}) \\ \text{subject to :} \quad & \mathbf{A}\mathbf{v} = \mathbf{0}, \\ & \mathbf{v}^{\text{LB}}(\mathbf{s}) \leq \mathbf{v} \leq \mathbf{v}^{\text{UB}}(\mathbf{s}), \end{aligned} \tag{2}$$

where $\mathbf{s}$ denotes the state variables describing the extracellular environment (e.g., concentrations of substrates, biomass, and products) with time derivative $\dot{\mathbf{s}}$ and initial conditions $\mathbf{s}_0$; $\mathbf{v}$ denotes the metabolic fluxes that include both intracellular fluxes and exchange rates; $\mathbf{A}$ is the stoichiometric matrix of the metabolic network; and $\mathbf{v}^{\text{LB}}(\mathbf{s})$ and $\mathbf{v}^{\text{UB}}(\mathbf{s})$ are the lower and upper bounds on the fluxes, respectively, which are functions of the extracellular concentrations. The vector function $\mathbf{f}$, specified by the set of mass balances in the extracellular medium, defines the rate of change of each component of $\mathbf{s}$ and must be integrated to determine the time evolution of extracellular concentrations. The scalar function h is the cellular objective that is maximized by the cells. Whenever more than one microbial species are present in the culture, then multiple flux balance models of the form (2) must be incorporated into (1) [10].

DFBA models can be classified as ordinary differential equations with embedded optimization wherein the lower-level FBA optimization can either be a linear or nonlinear program [30]. A variety of methods have been developed for integrating DFBA models, which are summarized in S1 Text. We focus on the direct approach for integrating DFBA models in this work due to its ability to ensure accurate solutions through the use of error-controlled integration schemes. Another advantage of the direct approach is that a unique solution set to the FBA (2) can be obtained using lexicographic optimization [10, 27], which may help overcome numerical challenges that can occur when using alternative DFBA simulators (e.g., see [31, Chapter 3]). Since the direct approach requires continuous monitoring and identification of any active set changes in (2), it constitutes a dynamic simulation with discrete events (i.e., a hybrid system). In the next section, we present the proposed nsPCE method that is capable of directly accounting for the hybrid nature of DFBA models.

Polynomial chaos expansions

Theoretical background. We consider a DFBA model with a set of M input parameters that are denoted by $\boldsymbol{x} = (x_1, \ldots, x_M)$. These parameters can appear in the initial conditions $\mathbf{s}_0$, rate of change function $\mathbf{f}$, cellular objective h, and/or the flux limits $\mathbf{v}^{\text{LB}}$, $\mathbf{v}^{\text{UB}}$. We look to develop a computationally cheap-to-evaluate representation of some output of the DFBA model referred to as the *model response*. The model response $y = \mathcal{M}(\boldsymbol{x})$ can be any chosen function of the states or fluxes that appear in (1) and (2) including, for example, metabolite concentrations, growth rate, or time-to-consumption of any metabolite. The model response function $\mathcal{M} : \mathbb{R}^M \to \mathbb{R}$ need not be known analytically, and can be approximated using a finite number of model evaluations. We focus on the scalar response case y for notational simplicity. However, the developed procedure can be easily applied separately to each component of a vector of responses $\boldsymbol{y}$.

Unless the parameters $\boldsymbol{x}$ are perfectly known, they must be treated as uncertain. Parameter uncertainty can generally be represented by a random vector $\boldsymbol{X}$ with some known probability density function (PDF). In this case, the model response also becomes a random variable with

some unknown PDF that is implicitly defined by

$$Y = \mathcal{M}(\boldsymbol{X}), \quad \boldsymbol{X} \sim f_{\boldsymbol{X}}, \tag{3}$$

where $\sim$ denotes "distributed as" and $f_{\boldsymbol{X}}$ denotes the PDF of uncertain parameters. Determining the distribution $f_{\boldsymbol{Y}}$ (or its statistical moments) of the model response represents the forward UQ problem that can be tackled in various ways, the majority of which require extensive sampling that is not feasible whenever $\mathcal{M}$ is a computationally expensive model. The polynomial chaos expansion (PCE) method addresses this problem by constructing a *surrogate model* that accurately approximates $\mathcal{M}$, but is significantly cheaper to evaluate. The PCE surrogate model can also be straightforwardly applied to other UQ tasks, as discussed later in the Results section. Provided that Y has finite variance, it can be represented with a PCE as follows [17]

$$Y = \mathcal{M}(\boldsymbol{X}) = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^M} a_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{X}), \tag{4}$$

where $a_{\boldsymbol{\alpha}} \in \mathbb{R}$ are coefficients of the expansion, $\Psi_{\boldsymbol{\alpha}} : \mathbb{R}^M \rightarrow \mathbb{R}$ are multivariate polynomials, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)$ is a multi-index that identifies the degree of the multivariate polynomials in each of the input parameters X_i, and $\mathbb{N} = \{1, 2, \ldots\}$ is the set of positive integers. The polynomial basis functions are required to be orthonormal with respect to the parameter distribution, such that they satisfy

$$\mathbb{E}\{\Psi_{\boldsymbol{\alpha}}(\boldsymbol{X})\Psi_{\boldsymbol{\beta}}(\boldsymbol{X})\} = \int_S \Psi_{\boldsymbol{\alpha}}(\boldsymbol{x})\Psi_{\boldsymbol{\beta}}(\boldsymbol{x}) f_{\boldsymbol{X}}(\boldsymbol{x}) d\boldsymbol{x} = \delta_{\boldsymbol{\alpha}\boldsymbol{\beta}}, \quad \forall \boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{N}^M, \tag{5}$$

where S is the support of the distribution of $\boldsymbol{X}$ and $\delta_{\boldsymbol{\alpha}\boldsymbol{\beta}}$ is the Kronecker delta that is 1 whenever $\boldsymbol{\alpha} = \boldsymbol{\beta}$ and 0 otherwise. For computational purposes, the series (4) must be truncated after a finite number of P terms, which yields the following approximation

$$Y^{PCE} = \mathcal{M}^{PCE}(\boldsymbol{X}) = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} a_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{X}) = \boldsymbol{a}^{\top} \boldsymbol{\Psi}(\boldsymbol{X}), \tag{6}$$

where $\mathcal{A}$ is a finite set of multi-indices with cardinality equal to P, $\boldsymbol{a} \in \mathbb{R}^P$ is a vector of the coefficients, and $\boldsymbol{\Psi} : \mathbb{R}^M \rightarrow \mathbb{R}^P$ is a vector containing all polynomial basis functions. The expansion coefficients are defined to be those that minimize the mean-square error (MSE) between the exact representation (4) and the truncated PCE (6)

$$\boldsymbol{a} = \underset{\tilde{\boldsymbol{a}} \in \mathbb{R}^P}{\operatorname{argmin}} \, \mathbb{E}\{(\mathcal{M}(\boldsymbol{X}) - \tilde{\boldsymbol{a}}^{\top} \boldsymbol{\Psi}(\boldsymbol{X}))^2\} = \mathbb{E}\{\mathcal{M}(\boldsymbol{X})\boldsymbol{\Psi}(\boldsymbol{X})\}. \tag{7}$$

The right-hand side of this expression represents the analytic solution to the MSE optimization problem and directly follows from the Hilbert projection theorem [32].

The expressions in (5) and (7) involve multivariate integration over complicated nonlinear functions. As such, the construction of the polynomial basis and computation of the expansion coefficients are usually carried out numerically in practice, which leads to additional sources of error. The choice of $\mathcal{A}$ also plays an important role in PCE performance because $\mathcal{A}$ directly controls the number of coefficients that must be estimated. Larger P values require more computational effort and are more susceptible to numerical sources of error. An overview of state-of-the-art methods for addressing these challenges is provided next.

Orthonormal basis construction. The complexity of determining the polynomials $\{\Psi_{\boldsymbol{\alpha}}(\boldsymbol{X})\}_{\boldsymbol{\alpha} \in \mathcal{A}}$ depends fully on the structure of the PDF $f_{\boldsymbol{X}}$. Whenever the uncertain parameters are statistically independent, then (5) reduces to the tensor product of M univariate polynomials that are orthonormal with respect to each marginal density f_{X_i}. These polynomials have

been analytically derived for many common PDFs [17], and can be found numerically for generic PDFs using algorithms in terms of the three-term recurrence relationship for orthogonal polynomials [33]. There are two main approaches for handling the more general case that $\boldsymbol{X}$ has statistically dependent (or correlated) elements. The first approach involves transforming the generic random vector $\boldsymbol{X}$ into a standard random vector $\boldsymbol{Z}$ for which it is simpler to build the polynomial basis functions [34]. Any *isoprobabilistic transformation* that preserves the PDFs of these random vectors can be utilized, though the most commonly used is the Rosenblatt transformation [35]. The second approach involves applying a more sophisticated numerical procedure that is able to impose the conditions in (5) simultaneously in M dimensions. This includes the Gram-Schmidt process [36] as well as the modified Cholesky decomposition of the Gram moment matrix [37, 38].

Sparse truncation and regression. We denote the approximate PCE with numerically estimated coefficients $\hat{\boldsymbol{a}}$ as follows

$$\hat{Y}^{PCE} = \hat{\mathcal{M}}^{PCE}(\boldsymbol{X}) = \hat{\boldsymbol{a}}^{\top}\boldsymbol{\Psi}(\boldsymbol{X}). \tag{8}$$

A variety of methods have been proposed for estimating the coefficients that can be broadly categorized as *intrusive* (e.g., Galerkin projection [12]) or *non-intrusive* (e.g., pseudo-spectral projection [39] or regression [15]). Here, we focus exclusively on non-intrusive methods. The phrase "non-intrusive" implies that coefficient estimates are obtained over a finite set of parameter realizations $\mathcal{X} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$, referred to as the experimental design (ED). These samples can be chosen in various ways including Monte Carlo sampling, quasi-random samples derived from Sobol or Halton sequences, or sparse grids to name a few [40]. The computational model is then evaluated at every point in the ED, i.e., $\mathcal{Y} = \{y^{(1)}, \ldots, y^{(N)}\}$ with $y^{(i)} = \mathcal{M}(\boldsymbol{x}^{(i)})$ for all $i = 1, \ldots, N$. As such, non-intrusive approaches are "black-box" in the sense that they can be applied to any function, even when this function is not explicitly known, and do not require any modification to the deterministic solver.

We will focus on regression methods due to their flexibility when it comes to enforcing sparsity. In the regression approach, coefficients $\hat{\boldsymbol{a}}$ are defined as those that minimize the least-square residual of the polynomial approximation over the ED $\mathcal{X}$

$$\hat{\boldsymbol{a}} = \underset{\tilde{\boldsymbol{a}} \in \mathbb{R}^P}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N}\left(\mathcal{M}(\boldsymbol{x}^{(i)}) - \tilde{\boldsymbol{a}}^{\top}\boldsymbol{\Psi}(\boldsymbol{x}^{(i)})\right)^2 = (\boldsymbol{A}^{\top}\boldsymbol{A})^{-1}\boldsymbol{A}^{\top}\mathcal{Y}, \tag{9}$$

where $\boldsymbol{A} \in \mathbb{R}^{N \times P}$ is the model matrix that contains the values of all polynomial basis functions evaluated at all ED points. The solution of (9) requires a minimum number of sample points $N \geq P$ to ensure a unique solution exists. Since every sample requires an expensive DFBA simulation here, the truncation scheme plays a central role in reducing the complexity of surrogate model construction. The total degree method is the most commonly used approach for specifying $\mathcal{A}$, which looks to keep all polynomials up to a specified order p in the series. For total degree truncation, the set of multi-indices is defined as $\mathcal{A} = \{\boldsymbol{\alpha} \in \mathbb{N}^M : \|\boldsymbol{\alpha}\|_1 \leq p\}$, where $\|\boldsymbol{\alpha}\|_1 = \alpha_1 + \cdots + \alpha_M$ and $P = \frac{(M+p)!}{M!p!}$. Due to the sharp increase in P as the polynomial order increases, the total degree truncation scheme can quickly lead to a prohibitive number of model evaluations, especially in high dimensions. This issue is often termed the *curse-of-dimensionality*, which is known to considerably limit standard PCE methods.

We look to take advantage of two approaches for overcoming the curse-of-dimensionality limitation. The first approach involves replacing the total order truncation with the so-called

hyperbolic truncation scheme, which is defined as

$$\mathcal{A}^{M,p,q} = \{\boldsymbol{\alpha} \in \mathbb{N}^M : \|\boldsymbol{\alpha}\|_q \leq p\}, \qquad \|\boldsymbol{\alpha}\|_q = \left(\sum_{i=1}^{M} \alpha_i^q\right)^{1/q}, \tag{10}$$

where $0 < q \leq 1$. Lower values for q limit the number of high-order interaction terms considered, which directly lead to sparser solutions. The second approach looks to further sparsify the solution, without sacrificing potentially important interaction terms, by including a regularization term of the form $\lambda\|\tilde{\boldsymbol{a}}\|_1$ with $\lambda \geq 0$ in the least-squares problem (9). This regularization term is known to force the minimization to favor low-rank solutions and ensures the existence of a unique solution even when $N < P$.

The key challenge with regularization is a proper choice of λ, which indirectly specifies the number of non-zero coefficients included in the expansion. In this work, we use the hybrid least angle regression (LAR) method to solve the regularized version of (9). LAR is an efficient procedure for variable selection, which aims to select the predictors (i.e., polynomials $\Psi_{\boldsymbol{\alpha}}$) that have the greatest impact on the model response among a potentially large set of candidates [41]. Hybrid LAR is a variant of the original LAR that uses a modified cross-validation scheme to estimate the approximation error [19]. This modification relies on only a single call to the LAR procedure, which provides significant savings in computational cost when compared to the original method. The relative MSE (RMSE), which is defined as $\varepsilon = \text{MSE}/\text{Var}\{Y\}$, is the natural choice of the approximation error in PCE and can be robustly estimated by the leave-one-out (LOO) cross-validation error ε_{LOO}. Not only can ε_{LOO} be calculated analytically for PCE models [42], but it is known to be much less sensitive to overfitting than the empirical estimator [43].

Provided a sensible sampling strategy has been chosen, the remaining parameters that must be selected are related to truncation p and q and the ED size N. We use a systematic procedure for selecting these parameters to achieve a target error level ε_{target}. As discussed in [19], a *basis-adaptive* strategy can help overcome potential limitations of an *a priori* fixed truncation set $\mathcal{A}$ by letting the maximum degree be driven directly from the data. The basic idea is to start with small values for p and q, estimate the coefficients using hybrid LAR, and calculate ε_{LOO}. These steps are repeated for incremented values of p and q, and the algorithm returns the PCE model with the lowest error. Early stop criteria can easily be introduced to avoid an excessive number of iterations. However, when dealing with computationally expensive models, the number of model evaluations N dominates the cost of construction of the surrogate model. We therefore propose an iterative "greedy" approach for constructing the ED to ensure that N can be kept as small as possible. This sequential ED strategy can be summarized as

1. Initialize the current ED with a relatively small number of samples N_{init}.
2. Train a sparse basis-adaptive PCE using the current ED and calculate ε_{LOO}.
3. If $\varepsilon_{LOO} < \varepsilon_{target}$, stop the algorithm and return current PCE. Otherwise, enrich the current ED with N_{add} more samples and return to Step 2.

Note that any method can be used in the training step of this algorithm. Thus, in the proposed nsPCE method, the desired accuracy level is the key parameter that must chosen by the user.

The nsPCE surrogate modeling method

The PCE method is guaranteed to converge as both the number of model evaluations N and number of terms in the expansion P increase; however, the rate of convergence can be very slow whenever $\mathcal{M}$ exhibits any singularities [24]. This is a primary challenge in DFBA models

since they can lose differentiability when a switch in the active set of the FBA problem (2) occurs. Inspired by [25], we look to take advantage of the following *multi-element* representation of PCE as it is capable of capturing non-smooth behavior

$$Y = \mathcal{M}(\boldsymbol{X}) = \sum_{k=1}^{N_e} \sum_{\boldsymbol{\alpha} \in \mathbb{N}^M} a_{k,\boldsymbol{\alpha}} \Psi_{k,\boldsymbol{\alpha}}(\boldsymbol{X}) I_{S_k}(\boldsymbol{X}), \tag{11}$$

where N_e denotes the number of elements; S_k, $a_{k,\boldsymbol{\alpha}}$, and $\Psi_{k,\boldsymbol{\alpha}}$ denote the local support, coefficient, and orthogonal polynomials in element k, respectively; $S = \bigcup_{k=1}^{N_e} S_k$; and $I_{S_k}(\boldsymbol{X})$ are indicator random variables defined by

$$I_{S_k}(\boldsymbol{X}) = \begin{cases} 1 & \text{if } \boldsymbol{X} \in S_k \\ 0 & \text{otherwise.} \end{cases} \quad k = 1, \ldots, N_e \tag{12}$$

The indicator random variables can be used to define the following conditional random variables $\boldsymbol{X}_k = \boldsymbol{X} | (I_{S_k}(\boldsymbol{X}) = 1)$ with PDF

$$f_{X_k}(\boldsymbol{x}_k) = \frac{f_X(\boldsymbol{x}_k)}{\Pr(I_{S_k}(\boldsymbol{X}) = 1)} = \frac{f_X(\boldsymbol{x}_k)}{\int_{S_k} f_X(\boldsymbol{x}) d\boldsymbol{x}}. \tag{13}$$

The local polynomials in (11) are orthogonal with respect to $\boldsymbol{X}_k$ while the coefficients are similarly defined as in (7) but now in terms of $\boldsymbol{X}_k$. This implies that the same strategies discussed above for building the polynomials, estimating the coefficients using regularized least squares, truncating the expansion, and sequentially populating the ED can be utilized locally within each element.

The remaining unanswered question is how to design the elements $\{S_k\}_{k=1}^{N_e}$ to limit the growth in the number of model evaluations since N will scale approximately linearly with N_e. The best decomposition should ensure that the model response behaves smoothly in every element. The proposed nsPCE method decomposes the support into two elements S_1 and S_2 that denote, respectively, the set of parameters for which the singularity has not and has occurred. This idea is best illustrated through a simple example. Consider the following non-smooth ODE system $\dot{y} = -x$ if $y > 0$ and $\dot{y} = 0$ otherwise with initial condition $y_0 > 0$, whose solution is given by

$$y(t, x) = \begin{cases} y_0 - tx, & \text{if } y_0 > tx, \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

This function is not differentiable at the time point $t_s(x) = y_0/x$, which can be thought of as the "singularity manifold" in the parameter support space, i.e., t_s is the boundary function that separates S_1 and S_2. At any given time of interest t, the two elements can be defined in terms of $t_s(x)$ as follows

$$S_1(t) = \{x \in S : t_s(x) > t\}, \quad S_2(t) = \{x \in S : t_s(x) \leq t\}. \tag{15}$$

Let us briefly analyze the behavior of these elements. The elements are continuous functions of time, meaning that every time of interest t requires a different decomposition. Whenever t is outside of the support of $t_s(X)$, then one of these sets is empty and we revert back to traditional PCE that covers the full support S. In light of this, we can easily generalize the idea to the case of multiple $n_s > 1$ sequential singularities as long as the random variables $\{t_{s_i}(X)\}_{i=1}^{n_s}$ do not have overlapping supports. When multiple non-overlapping singularities are present, we must

simply find the support in which t lies and define the two elements using that corresponding boundary function. The case of overlapping supports is more challenging due to the fact that more elements would need to be created based on the intersection of S_1 and S_2 for all active singularities.

For the simple scalar example in (14), we can analytically derive the boundary function; however, this is not generally possible in DFBA models. Based on the observation that the singularity boundary depends smoothly on the parameters, we instead propose to construct a sparse PCE model to approximate the boundary in multiple dimensions, i.e., $t_s \approx \hat{t}_s^{PCE}$. The nsPCE method thus creates a surrogate model with the following structure for any $\boldsymbol{x} \in S$

$$\hat{\mathcal{M}}^{nsPCE}(\boldsymbol{x}) = \sum_{k=1}^{2} \hat{\boldsymbol{a}}_k^{\top} \boldsymbol{\Psi}_k(\boldsymbol{x}) I_{S_k}(\boldsymbol{x}) = \begin{cases} \hat{\boldsymbol{a}}_1^{\top} \boldsymbol{\Psi}_1(\boldsymbol{x}), & \text{if } \boldsymbol{x} \in S_1, \\ \hat{\boldsymbol{a}}_2^{\top} \boldsymbol{\Psi}_2(\boldsymbol{x}), & \text{if } \boldsymbol{x} \in S_2, \end{cases} \tag{16}$$

where the coefficients $\hat{\boldsymbol{a}}_k$ are estimated from the sparse regression problem

$$\hat{\boldsymbol{a}}_k = \underset{\tilde{\boldsymbol{a}}_k}{\operatorname{argmin}} \ \frac{1}{N_k} \|\mathcal{Y}_k - \boldsymbol{A}_k \tilde{\boldsymbol{a}}_k\|_2^2 + \lambda_k \|\tilde{\boldsymbol{a}}_k\|_1, \qquad k \in \{1, 2\} \tag{17}$$

based on the local ED $\mathcal{X}_k = \{\boldsymbol{x}_k^{(1)}, \ldots, \boldsymbol{x}_k^{(N_k)}\}$ and $\mathcal{Y}_k = \{y_k^{(1)}, \ldots, y_k^{(N_k)}\}$ in terms of N_k samples. Notice that the full DFBA model must be integrated when constructing $\hat{t}_s^{PCE}$. Instead of discarding this information, it can be reused by storing the list of state and time points generated when integrating the DFBA model and then interpolating these points when calculating the model response function. Thus, we can use this approach to initialize the ED $\mathcal{X}$, model response data $\mathcal{Y}$, and singularity time data $\mathcal{T}_s$. Using $\mathcal{T}_s$ along with the set definitions in (15), we can easily partition $\mathcal{X}$ and $\mathcal{Y}$ into the required local EDs. The sequential ED strategy is then applied in each element to ensure that the target error is achieved.

A flowchart that summarizes the main steps of the nsPCE method is shown in Fig 1. By evaluating the nsPCE surrogate in (16), which is much cheaper to evaluate than the full model $\mathcal{M}$, on a collection of Monte Carlo samples of the parameters, we can directly approximate statistical properties of Y including moments, parametric sensitivities, or even its full distribution.

Numerical implementation

The complete set of Matlab scripts that implement the nsPCE method is available at [29]. All of the modifiable parameters in the algorithm are defined in the "User inputs" section of the `main_pce.m` script, which automatically executes the steps summarized in Fig 1. It is important to note that the scripts require the installation of two additional packages that integrate the DFBA model and construct sparse PCE models. The nsPCE scripts are written to be compatible with readily available DFBA and PCE toolboxes to provide flexibility. The simulation of DFBA models can be done with any non-smooth integration code including `COBRA` [44], `ORCA` [45], and `DFBAlab` [10]. All files needed by the DFBA integrator should be placed in the `dfba_model` folder. We opt for `DFBAlab` in this work due to certain numerical advantages that it exhibits over the available alternatives (see [27, 31] for more details). The sparse PCE operations are carried out using `UQLab` [43], which implements the hybrid LAR method as well as the required calculations to determine the cross-validation error ε_{LOO}. The syntax in `main_pce.m` is heavily based on `UQLab`. Hence, some modifications to the source code may be needed to perform the same operations with other toolboxes.

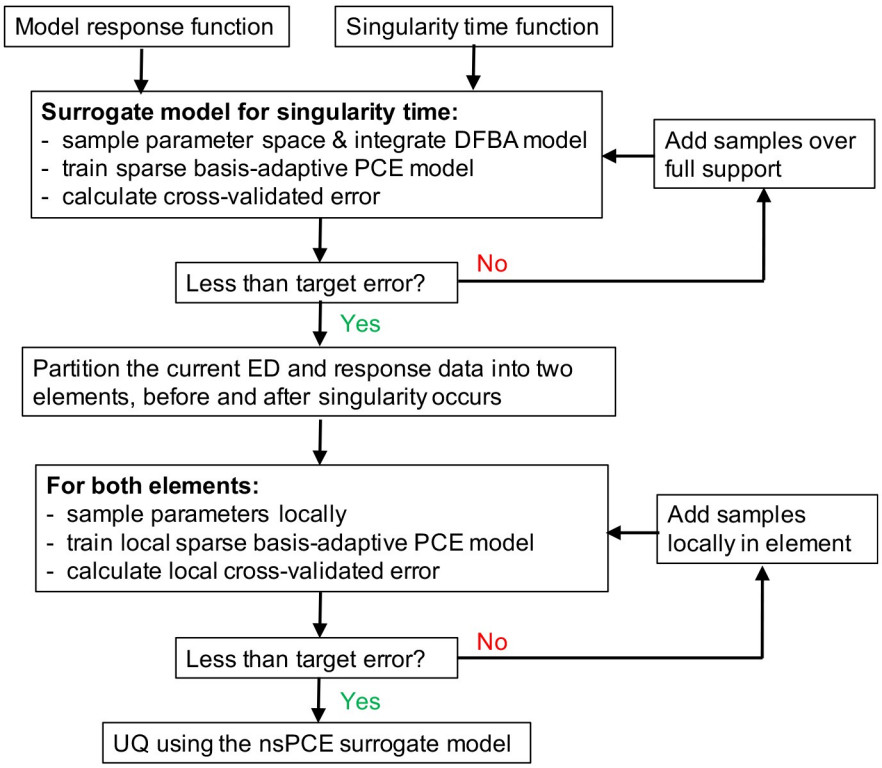

Fig 1. Flowchart for the proposed nsPCE surrogate modeling method. The model response function can be freely chosen by the user. The singularity time function should be specified implicitly as a function of the DFBA model states. This function can be identified by simulating the DFBA model with nominal parameters and locating at which time points a switch in the active set of the FBA solution occurs. The PCE coefficients are fit using the basis-adaptive version of the hybrid LAR method, while the ED is sequentially enriched to ensure that the target accuracy level is achieved.

https://doi.org/10.1371/journal.pcbi.1007308.g001

Results

We present two separate case studies in this section. The first case study explores Bayesian estimation of six parameters related to the substrate uptake kinetics in a computationally expensive DFBA model of *E. coli* with a genome-scale metabolic network. The goal of the first case study is to demonstrate advantages of the proposed nsPCE method over alternatives as well as its application to a realistic problem that has been previously studied in the literature. The second case study focuses on maximum a posteriori estimation in a synthetic DFBA problem with a relatively large number of parameters, i.e., twenty uncertain parameters appearing in a variety of intracellular and extracellular quantities. The goal of the second case study is to provide preliminary evidence of the scalability of nsPCE as well as the fact that the method is applicable to a wide-variety of UQ applications.

Case study 1: Batch fermentation of *E. coli* monoculture

This case study is based on a DFBA model of a batch fermentation reactor consisting of an *E. coli* monoculture, which has been investigated for the production of valuable chemicals such as ethanol. Here, we focus on the initial phase of batch operation of the *E. coli* fermentation reactor under aerobic growth in a glucose and xylose mixed media [8]. No ethanol production is observed under aerobic conditions (i.e., this phase is mainly used to increase the biomass), such that the concentration of ethanol can be omitted from the dynamics. This case study is

commonly used as a benchmark for comparing DFBA solvers (see, e.g., [16, 27, 31]), as it exhibits stiff dynamics and multiple singularities.

The dynamic mass balance equations of the form (1) for the extracelluar environment can be summarized as follows

$$\begin{aligned} \dot{b}(t) &= \mu(t)b(t), \\ \dot{g}(t) &= -u_g(t)b(t), \\ \dot{z}(t) &= -u_z(t)b(t), \end{aligned} \tag{18}$$

where $b(t)$, $g(t)$, and $z(t)$ denote the biomass, glucose, and xlyose concentrations at time t, respectively. The uptake kinetics for glucose, xylose, and oxygen are given by Michaelis-Menten kinetics

$$\begin{aligned} u_g(t) &= u_{g,max}\frac{g(t)}{K_g + g(t)}, \\ u_z(t) &= u_{z,max}\frac{z(t)}{K_z + z(t)}\frac{1}{1+\frac{g(t)}{K_{ig}}}, \\ u_o(t) &= u_{o,max}\frac{o(t)}{K_o + o(t)}, \end{aligned} \tag{19}$$

where parameters $u_{g,max}$, $u_{z,max}$, $u_{o,max}$, K_g, K_z, K_o, and K_{ig} correspond to the maximum substrate uptake rates, saturation constants, and inhibition constants. It is assumed that the reactor oxygen concentration, $o(t)$, is controlled and is therefore constant. The growth rate $\mu(t)$, on the other hand, is determined from the metabolic network model of wild-type *E. coli*. The chosen metabolic network reconstruction was iJR904 [28], which contains 1075 reactions and 761 metabolites. The cells are assumed to maximize growth, implying (2) is an LP of the form

$$\begin{aligned} \mu(t) = \min_{\mathbf{v}} \quad & \mathbf{c}^\top \mathbf{v}, \\ \text{s.t.} \quad & \mathbf{A}\mathbf{v} = \mathbf{0}, \\ & v_{g_{ext}} = u_g(t), \\ & v_{z_{ext}} = u_z(t), \\ & v_{o_{ext}} = u_o(t), \\ & \mathbf{v}^{\mathrm{LB}} \le \mathbf{v} \le \mathbf{v}^{\mathrm{UB}}, \end{aligned} \tag{20}$$

where $\mathbf{c}$ is a vector of weights that represent the contribution of each flux to biomass formation while $v_{g_{ext}}$, $v_{z_{ext}}$, and $v_{o_{ext}}$ are, respectively, the exchange fluxes for glucose, xylose, and oxygen (i.e., elements of the flux vector $\mathbf{v}$). Thus, the metabolic network interacts with the extracellular environment through the exchange fluxes in (19).

The initial conditions of the batch are assumed to be fixed at 0.03 g/L of inoculum, 15.5 g/L of glucose, and 8 g/L of xylose; the oxygen concentration is kept constant at 0.24 mmol/L; and $\mathbf{A}$, $\mathbf{c}$, $\mathbf{v}^{\mathrm{LB}}$, and $\mathbf{v}^{\mathrm{UB}}$ are specified by the iJR904 model. However, the parameters in the substrate uptake rates (19) should be fit to experimental data since they cannot be easily predicted from first principles. This problem of identifying the model parameters was partially tackled in [8], where most of the parameters were fixed according to estimates provided in the literature

while $u_{z,max}$ and K_{ig} were adjusted by trial-and-error to match transient measurements of biomass, glucose, and xylose. The reported parameter estimates are given in S1 Table. Since $o(t)$ is fixed, $u_{o,max}$ and K_o can be lumped into a single parameter u_o. These six parameters are unknown and here are modeled as a random vector whose elements are independent and uniformly distributed around ±10% of the nominal values. We selected this range to reflect a reasonable level of confidence in the reported literature values. In the following, we demonstrate how the proposed nsPCE surrogate modeling method can facilitate UQ tasks that are otherwise computationally intractable with respect to the full DFBA model.

All reported computations are performed in MATLAB R2016a on a MacBook Pro with 8 GB of RAM and a 2.6 GHz Intel i5 processor. The DFBA model is simulated using DFBAlab with default options for integration and LP optimization tolerances. CPLEX was used as the LP solver and MATLAB ode15s was used as the integrator.

Decomposition of parameter space. Before selecting the element decomposition, we must first simulate the DFBA model to locate any significant singularities. The extracellular glucose, xylose, and biomass concentration profiles are plotted in Fig 2 for one hundred randomly sampled parameter values. For a given realization of the parameter, the full simulation requires approximately 1.5 seconds of CPU time.

At the start of the batch, glucose is consumed preferentially over xylose. Once glucose has been depleted, the LP solution switches and xylose becomes the main carbon source. The final batch time is then specified as the time that both glucose and xylose have been fully depleted, at which point the LP becomes infeasible and the solution ceases to exist. The *E. coli* cells stop growing at this point due to the lack of a carbon source. Although physically the cells would begin to die in this situation, DFBA models cannot directly predict the cell death phase and thus we assume the biomass remains constant for simplicity. The time-to-consumption of

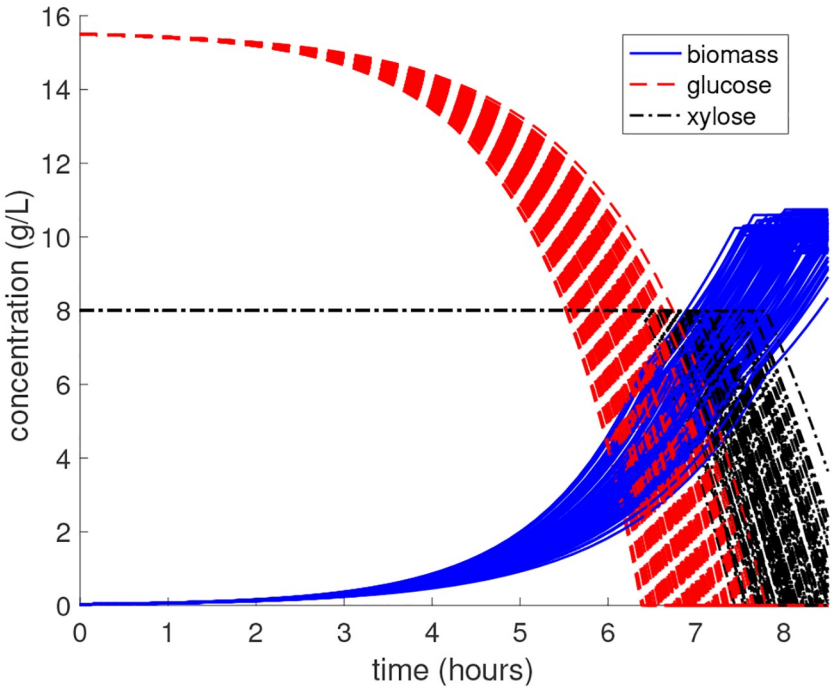

Fig 2. Monte Carlo simulation of *E. coli* DFBA model. The genome-scale model is integrated from 0 to 8.5 hours for 100 different parameter realizations that are independently drawn from the uniform prior density. The consumption of xylose only occurs after glucose is fully exhausted, which is a strong function of the parameters.

https://doi.org/10.1371/journal.pcbi.1007308.g002

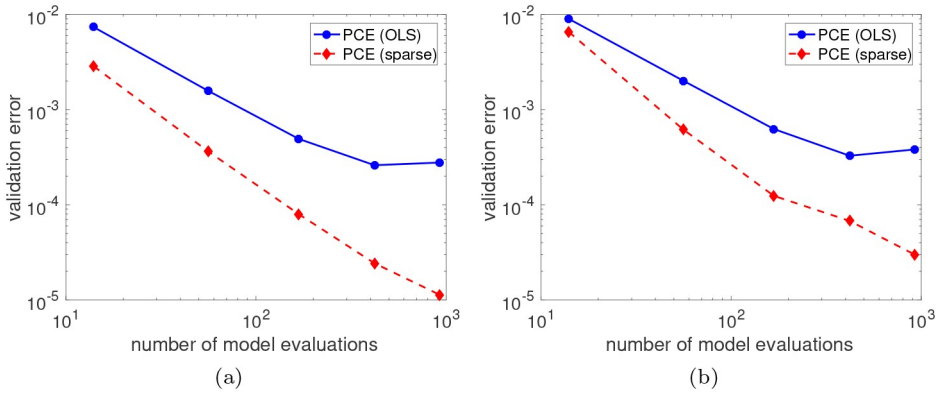

Fig 3. Accuracy of singularity time surrogate models. RMSE versus the number of model evaluations (i.e., size of the experimental design) used to train the PCE model for **(a)** the glucose singularity t_g and **(b)** the xylose singularity t_z. The RMSE was estimated empirically from a validation set of 10,000 full DFBA simulations.

https://doi.org/10.1371/journal.pcbi.1007308.g003

glucose t_g and xylose t_z represent the two singularities in this problem, and clearly depend on the value of the model parameters. Since the singularity time functions cannot be derived analytically, we look to construct PCE approximations for both t_g and t_z. We investigate two different fitting methods: classical *full* PCE with coefficients estimated using ordinary least squares (OLS) and *sparse* basis-adaptive PCE with coefficients estimated using hybrid LAR. The degree of the polynomials is varied from 1 to 6 in the full PCE method, where $N = 2P$ model evaluations are used for regression with P denoting the size of the basis. In the sparse PCE method, the maximum degree is allowed to vary from 1 to 20, and a hyperbolic truncation scheme (10) is used with $q = 0.75$. The experimental designs (EDs) are generated using Monte Carlo (MC) sampling with a fixed random seed to ensure repeatable results. Fig 3a and 3b show the RMSE as a function of the number of model evaluations used to fit surrogate models for t_g and t_z, respectively. The sparse PCE method consistently outperforms full PCE, achieving approximately an order-of-magnitude lower RMSE for all ED sizes.

The sparse PCE surrogate models for t_g and t_z are used in the nsPCE method to build surrogates for the extracellular concentrations. Additionally, these surrogate models contain useful information on which parameters influence the consumption of different substrates. The Sobol' indices of $t_g(\boldsymbol{X})$ and $t_z(\boldsymbol{X})$ are shown in Fig 4, which are a commonly used tool in global sensitivity analysis for ranking the parameters according to their contribution to the variance of the model response. The Sobol' indices can be computed *analytically* from the PCE coefficients [43], which requires less than one second of CPU time here. It is interesting to note that $u_{g,max}$ and u_o mainly contribute to the variance of $t_g(\boldsymbol{X})$, while $u_{g,max}$, $u_{z,max}$, and u_o are the significant contributors to the variance of $t_z(\boldsymbol{X})$.

The surrogate models can also be used to estimate the PDF of $t_g(\boldsymbol{X})$ and $t_z(\boldsymbol{X})$, as shown in Fig 4. From the estimated PDFs, we find that $t_g(\boldsymbol{X})$ ranges from approximately 6.31 to 7.87 hr, whereas $t_z(\boldsymbol{X})$ ranges from approximately 7.33 to 9.12 hr. This suggests that the model response is a non-smooth function of $\boldsymbol{X} \in S$ for any $t \in [6.31, 9.12]$ hr, so that we must split S into two disjoint regions according to 15. Since the supports of $t_g(\boldsymbol{X})$ and $t_z(\boldsymbol{X})$ partially overlap for any $t \in [7.33, 7.87]$ hr, additional elements should be introduced to ensure the model response is smooth. However, for times outside of this window, we can exclusively define the elements of the parameter space in terms of t_g for times before 7.33 hr and t_z for times after 7.87 hr. Plots of these two regions at times 6.5, 7.0, and 7.25 hr projected onto the two most sensitive

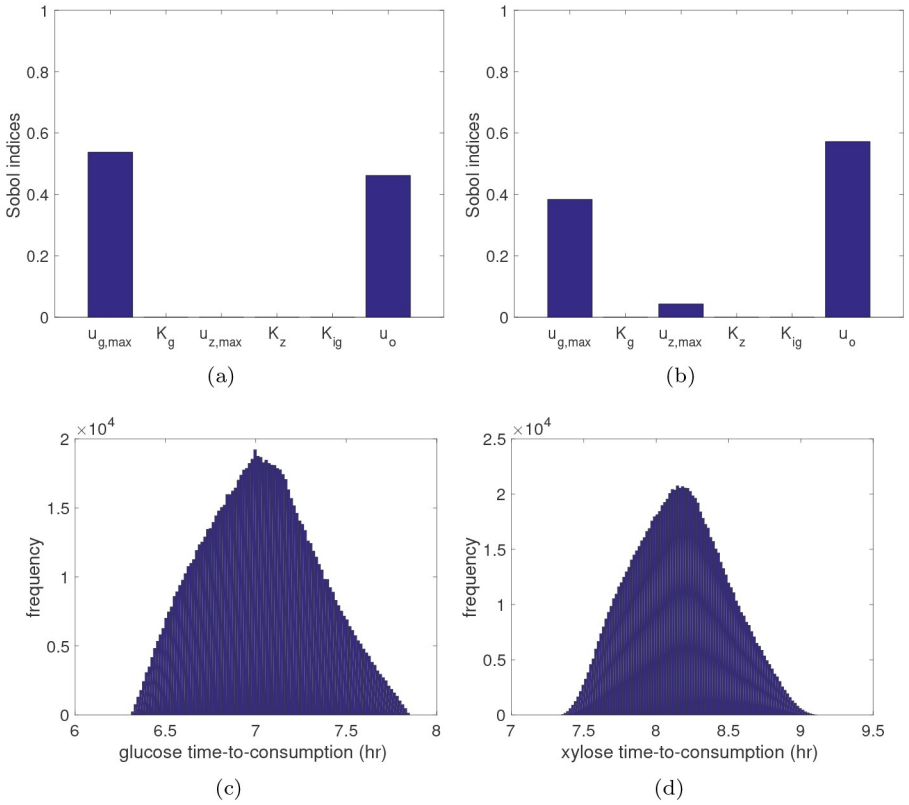

Fig 4. Uncertainty propagation with singularity time surrogate models. The estimated global sensitivity indices of **(a)** t_g and **(b)** t_z with respect to the uncertain parameters. The estimated PDF of **(c)** t_g and **(d)** t_z based on 1e+6 surrogate model evaluations, which only requires approximately 1 second of CPU time.

https://doi.org/10.1371/journal.pcbi.1007308.g004

parameters are shown in Fig 5. The blue region represents S_1 while the red region represents S_2. For comparison purposes, we also show the decision boundary in green (along with 95% confidence limits with dashed green lines) learned from a support vector machine (SVM) binary classifier [46] that was trained using the same 500 data points. The SVM model is unable to capture the significant nonlinear behavior of the boundary as it evolves over time. Thus, SVM results in relatively large misclassification errors due to the limited training data. The sparse PCE model, however, is able to accurately represent the t_g function over the full support (see the parity plot in Fig 5d), which leads to a much more accurate representation of these two elements using limited data.

The "true" RMSE values reported in Fig 3 were estimated using a large validation set that consisted of 10,000 evaluations of the full DFBA model, which required over 3 hours of CPU time. Ideally, these additional model evaluations could be avoided by directly estimating the RMSE from the ED either empirically or using cross-validation techniques. The empirical estimate of the RMSE is based on sample-based approximations to the integral expressions for mean and variance. Cross-validation obtains a more robust RMSE estimate by splitting the ED into various training and validation sets, fitting different models with each training set, and averaging the prediction error of each model. We focus exclusively on ε_{LOO} in this work. Table 1 gives the estimated RMSE values for the sparse PCE surrogate models fit using different ED sizes. We observe that the empirical estimator greatly underpredicts the "true" RMSE found from the large validation set. In fact, for the smallest size $N = 10$, the empirical estimate

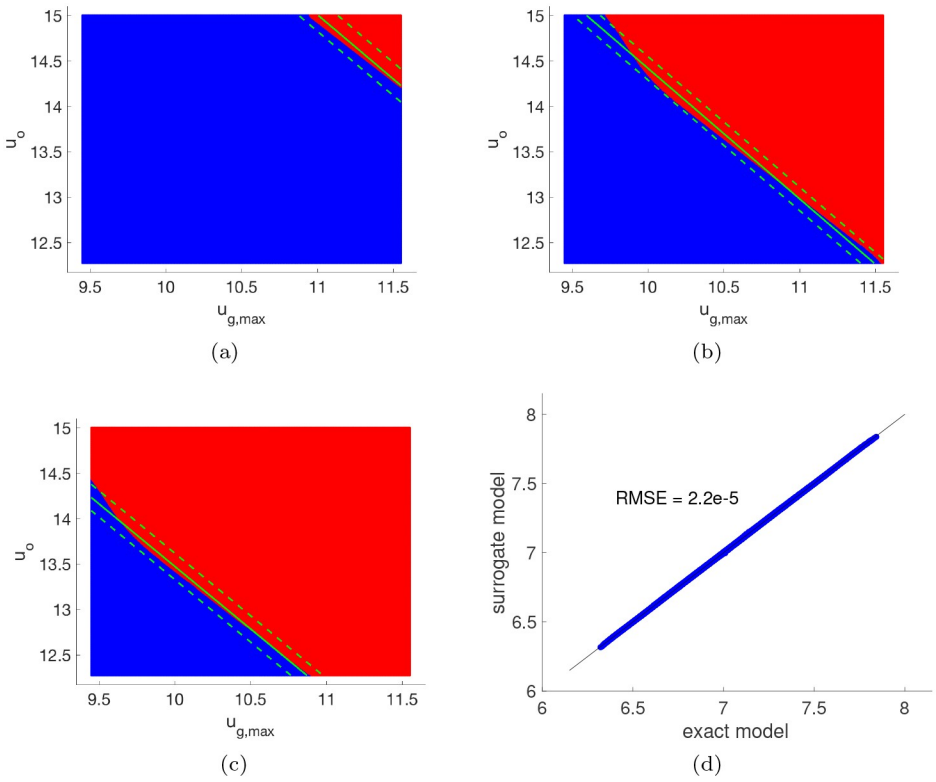

Fig 5. Parameter space decomposition over time. The decomposition of the parameter support into two non-overlapping elements at **(a)** 6.5 hr, **(b)** 7.0 hr, and **(c)** 7.25 hr using a sparse PCE model of the glucose singularity time t_g. The blue and red regions represent parameters for which $t_g(\boldsymbol{x}) > t$ and $t_g(\boldsymbol{x}) \leq t$, respectively, projected onto the two most sensitive parameters. The green line represents the decision boundary learned using an SVM classifier trained with the same set of data as the sparse PCE model, while the dashed green lines represent the corresponding 95% confidence limits. **(d)** Parity plot for the sparse PCE model of t_g for 1e4 validation points.

https://doi.org/10.1371/journal.pcbi.1007308.g005

is a factor of 10^4 smaller than the true RMSE. The cross-validated RMSE, on the other hand, predicts the correct order in all considered cases except $N = 10$ where it is off merely by a factor of 10 instead of 10^4. Note that ε_{LOO} is used within the hybrid LAR algorithm to select the best surrogate out of all potential candidates.

Table 1. Relative mean square error estimates for glucose singularity time surrogate models under multiple experimental design sizes.

N	Validation	Cross-validation	Empirical
10	1.014e-02	1.268e-03	2.601e-06
50	2.230e-04	3.718e-04	2.130e-04
100	1.616e-04	1.347e-04	5.333e-05
150	7.864e-05	6.468e-05	2.820e-05
200	5.787e-05	3.898e-05	2.416e-05
500	1.817e-05	1.273e-05	7.679e-06

The validation error is computed using a large set of samples not used in the fitting procedure. Cross-validation and empirical error, however, are computed using only points in the original experimental design. Cross-validation partitions the experimental design into various training and validation sets such that multiple models can be fit and their prediction errors averaged in order to compute more robust error estimates than its empirical counterpart. Here, a leave-one-out cross-validation procedure is utilized.

https://doi.org/10.1371/journal.pcbi.1007308.t001

Validation of nsPCE surrogate models. We have verified that the PCE surrogate models are able to accurately represent the singularity manifold that leads to non-smooth behavior in the states of the DFBA model. Thus, they can be used to build nsPCE surrogates for the extracellular concentrations based on the algorithm summarized in Fig 1. We choose three quantities of interest for illustrative purposes: glucose at time 7.0 hr, xylose at time 8.0 hr, and biomass at time 8.0 hr. We look to compare so-called global PCE to the proposed nsPCE method for these three quantities of interest. In global PCE, a single surrogate model is constructed over the full parameter support, while nsPCE systematically breaks down the support into two disjoint elements using the singularity time function as a dividing boundary. To ensure a fair comparison, the expansion coefficients of both global PCE and nsPCE are estimated using the basis-adaptive hybrid LAR strategy with maximum degree varying from 1 to 20 and $q = 0.75$ in the hyperbolic truncation scheme (10). In addition, the ED in both approaches are sequentially enriched using MC sampling with a fixed random seed. To simplify the construction of the polynomial basis functions when training the nsPCE surrogate models, the elements S_1 and S_2 were outerbounded with hyper-rectangles. However, only parameter values that explicitly fall within these sets are incorporated into the local ED. This simple approach for dealing with elements of any shape is currently used in the provided scripts [29], but other ways of dealing with generic elements can also be explored.

The convergence properties of the nsPCE surrogate models for the three quantities of interest are compared to that of global PCE in Fig 6. The nsPCE surrogates achieve significantly lower RMSE than the global PCE surrogates in virtually all cases considered, while requiring many fewer samples to converge to the target error level. In addition, global PCE saturates at the maximum number of ED samples for all three quantities of interest. This implies that global PCE is unable to achieve the desired accuracy levels, whereas nsPCE only saturates for the lowest target error of xylose. This behavior is expected since the convergence rate of global PCE is known to be substantially lowered whenever singularities are present in the model response function. Thus, nsPCE is able to significantly improve the rate of convergence based on a properly chosen elemental decomposition of the parameter support. To show that lower target error levels translate to improved predictions, parity plots for the three quantities of interest are shown in Fig 7. Note that global PCE has large prediction errors for particular values of the parameters (see the blue dots that largely deviate from the $y = x$ line), which is likely due to the fact that an inherently non-smooth function is being represented by smooth polynomials. This is highly undesirable when using the PCE to predict *specific* response values, as opposed to predicting statistical quantities that average over the response values where individual points are not as important. The nsPCE surrogate models clearly mitigate this limitation of global PCE in a significant way since there are no outlier predictions in the set of 10,000 validation points.

Bayesian parameter inference. Here, we focus on the inverse UQ problem of estimating parameters from data, which can be greatly accelerated using nsPCE. The same data set used in [8] is utilized, which includes measurements of the extracellular biomass, glucose, and xylose concentrations at $t \in \{5.5, 6.0, 6.5, 7.0, 7.25, 8.0, 8.25, 8.5\}$ hr. The measurements are corrupted with noise

$$\begin{aligned} D_i^b &= b(t_i; \boldsymbol{X}) + E_i^b, \qquad i = 1, \ldots, 8, \\ D_i^g &= g(t_i; \boldsymbol{X}) + E_i^g, \qquad i = 1, \ldots, 8, \\ D_i^z &= z(t_i; \boldsymbol{X}) + E_i^x, \qquad i = 1, \ldots, 8, \end{aligned} \tag{21}$$

where $\boldsymbol{D}_i = (D_i^b, D_i^g, D_i^z)$ and $\boldsymbol{E}_i = (E_i^b, E_i^g, E_i^z)$ are, respectively, vectors of the measured data

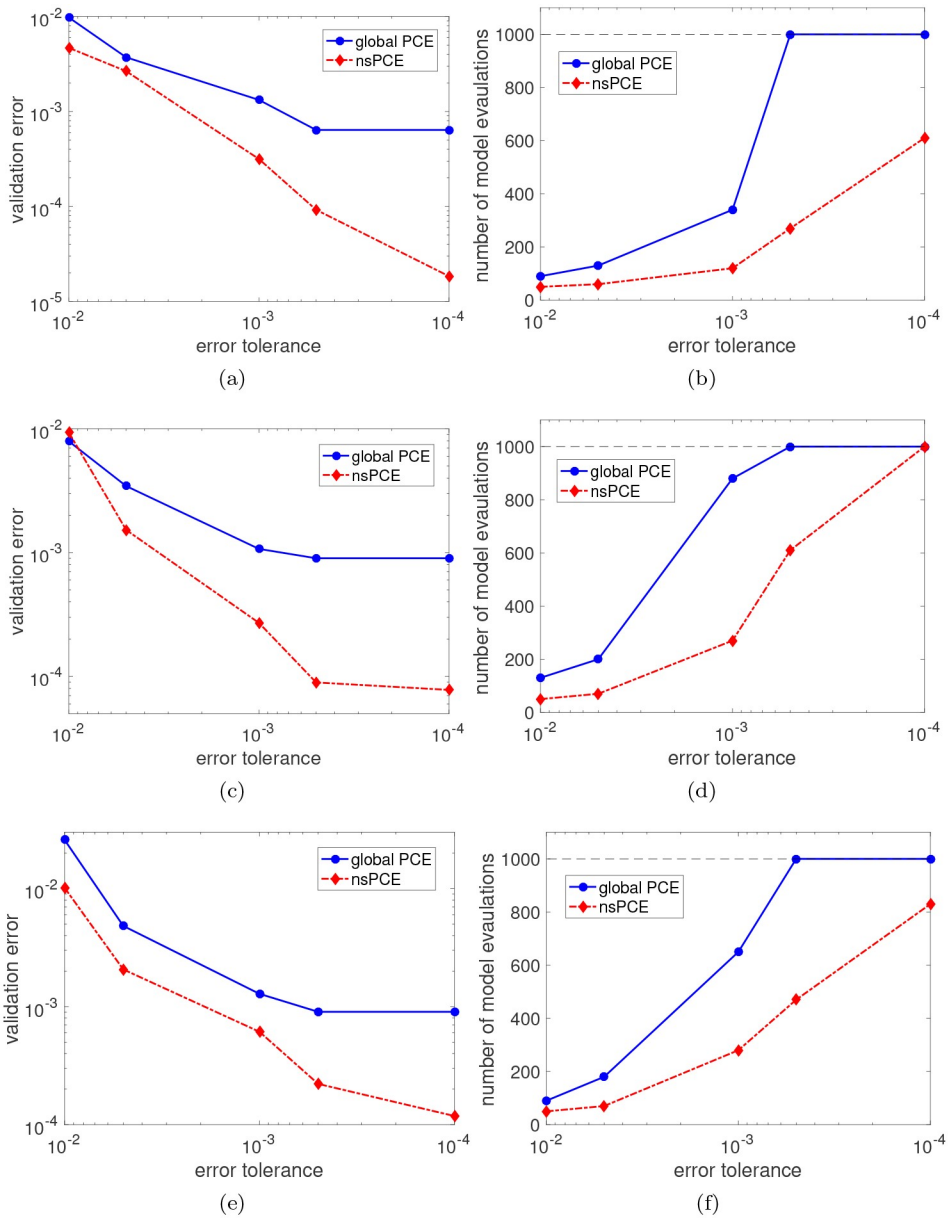

Fig 6. Convergence properties of nsPCE surrogate models. (a,b) Glucose concentration at time 7 hours. **(c,d)** Xylose concentration at time 8 hours. **(e,f)** Biomass concentration at time 8 hours. Left plots show the validation RMSE versus the specified error tolerance. Right plots show the total number of model evaluations based on a sequential ED construction, with a maximum of 1000 samples allowed. The global sparse basis-adaptive PCE results are also shown for comparison purposes.

https://doi.org/10.1371/journal.pcbi.1007308.g006

and noise at the ith time point. The concatenated data (respectively noise) vector is denoted by $\boldsymbol{D} = (\boldsymbol{D}_1, \ldots, \boldsymbol{D}_8)$ (respectively $\boldsymbol{E} = (\boldsymbol{E}_1, \ldots, \boldsymbol{E}_8)$). The measurement noise variables are modeled as independent zero-mean Gaussian random variables with state-dependent variance that equals 5% of the measured signal, i.e.,

$$E_i^v \sim \mathcal{N}(0, \sigma_{v,i}^2(\boldsymbol{X})), \quad \sigma_{v,i}(\boldsymbol{X}) = 0.05|v(t_i; \boldsymbol{X})|, \quad v \in \{b, g, z\}. \tag{22}$$

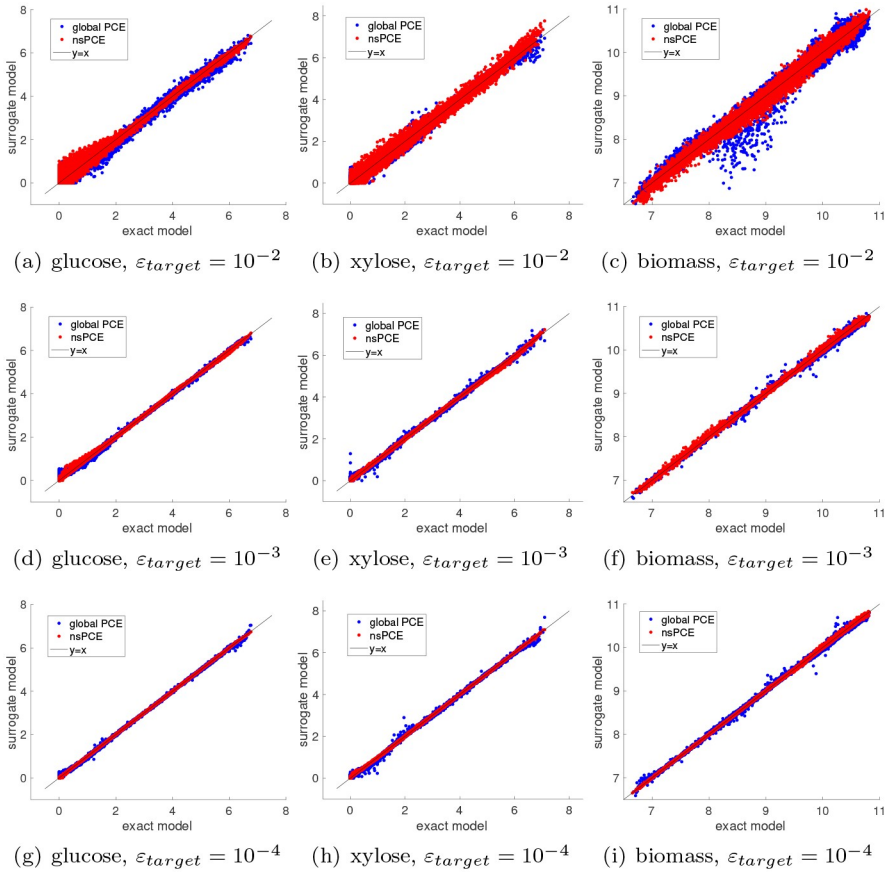

Fig 7. Parity plots for nsPCE surrogate models. (a,b,c) Target RMSE level $\varepsilon_{target} = 10^{-2}$. **(d,e,f)** Target RMSE level $\varepsilon_{target} = 10^{-3}$. **(g,h,i)** Target RMSE level $\varepsilon_{target} = 10^{-4}$. The left, middle, and right columns correspond to glucose concentration at 7 hours, xylose concentration at 8 hours, and biomass concentration at 8 hours, respectively. The parity plots for global sparse basis-adaptive PCE are overlaid for comparison purposes. The global PCE has considerably larger error than nsPCE.

https://doi.org/10.1371/journal.pcbi.1007308.g007

Given a set of measurements, the change in the state of information about the parameters is given by Bayes' rule [11]

$$f_{X|D}(\boldsymbol{x}|\boldsymbol{d}) = \frac{f_{D|X}(\boldsymbol{d}|\boldsymbol{x}) f_X(\boldsymbol{x})}{f_D(\boldsymbol{d})}, \tag{23}$$

where $f_{X|D}$ is the posterior density; $f_{D|X}$ is the likelihood function; f_X is the prior density; and f_D is the evidence. As Bayesian inference looks to characterize the full posterior density, it directly provides an explicit representation of the uncertainty in the parameter estimates.

The prior and likelihood function must be specified before solving (23). We assume the same uniform priors as those used to construct the nsPCE surrogate models, though these can differ in general. The likelihood function describes the discrepancy between the observed data and the model predictions in a probabilistic way. The likelihood function is specified by the data and noise models in (21) and (22), and is given by

$$f_{D|X}(\boldsymbol{d}|\boldsymbol{x}) = \prod_{i=1}^{8} \prod_{v \in \{b,g,z\}} \frac{1}{\sqrt{2\pi\sigma_{v,i}^2(\boldsymbol{x})}} \exp\left(-\frac{(d_i^v - v(t_i;\boldsymbol{x}))^2}{2\sigma_{v,i}^2(\boldsymbol{x})}\right). \tag{24}$$

Although we use a Gaussian likelihood here, the same Bayesian estimation approach can be applied to any choice of likelihood function and thus can be easily modified to incorporate other potentially important factors including sensor bias or asymmetric noise.

Since (23) cannot be solved analytically, we must resort to sample-based approximations that rely on generating samples from the target posterior distribution. A variety of methods have been developed for sampling from the unknown posterior $f_{\boldsymbol{X}|\boldsymbol{D}}$, including Markov Chain Monte Carlo (MCMC) [47–49] and sequential Monte Carlo (SMC) [50–52] algorithms. The proposed surrogate models can be used to accelerate any sampling-based method; however, we focus on SMC since this is a class of algorithms that can be made fully parallelized. SMC is based on the concept of *importance sampling*, which can be implemented in an iterative fashion such that the posterior is updated every time a new measurement becomes available. For a given number of particles N_p, the SMC approximation to (23) can be summarized as follows:

1. Initialization: set $k = 1$ and generate samples and weights $\{\boldsymbol{x}_i, w_i\}_{i=1}^{N_p}$ from prior.
2. Reweighting: update the weights $w_i \leftarrow w_i \times w_k(\boldsymbol{x}_i)$ where $w_k(\boldsymbol{x}_i) \propto f_{D_k|\boldsymbol{X}}(d_k|\boldsymbol{x}_i)$.
3. Resampling: resample $\{\boldsymbol{x}_i, w_i\}_{i=1}^{N_p}$ for particles with equal weights $\{\boldsymbol{x}_i^r, \frac{1}{N_p}\}_{i=1}^{N_p}$.
4. Loop: set $k \leftarrow k + 1$ and $\{\boldsymbol{x}_i, w_i\}_{i=1}^{N_p} \leftarrow \{\boldsymbol{x}_i^r, \frac{1}{N_p}\}_{i=1}^{N_p}$. Return to Step 2 if $k < k_f$.

When the algorithm stops at time k_f, the set of N_p particles targets the posterior distribution of interest. We use systematic resampling in Step 3 due to its computational simplicity and good empirical performance, though a variety of other methods are available [50]. Step 2 is usually the computational bottleneck because the model must be repeatedly solved in order to evaluate the likelihood weight factors using (24). Therefore, we propose to replace the evaluation of $v(t_i; \boldsymbol{x})$ with a nsPCE surrogate model $v^{nsPCE}(t_i; \boldsymbol{x})$ for every $v \in \{b, g, z\}$ and $i = 1, \ldots, 8$. We must then construct a total of 24 surrogates before running the SMC algorithm.

The same basic strategy described in the previous section is used for constructing all 24 of the nsPCE surrogate models. Similarly to how the samples for the singularity time are used to initialize the ED in each element, we can store the list of state and time points generated when integrating the DFBA model and interpolate these points to calculate the extracellular concentrations at every time point of interest. By keeping a working ED that is used to initialize each element at every time point, we can greatly limit the number of expensive DFBA simulations that represent the computational bottleneck in SMC. The proposed algorithm in Fig 1 is run with a target error of $\varepsilon_{target} = 10^{-3}$, 250 initial ED samples, 10 ED samples added at each iteration, 2500 maximum ED samples, maximum degree varying from 1 to 20, and hyperbolic truncation with $q = 0.75$. The algorithm converged with cross-validated errors ε_{LOO} below the desired tolerance using only a total of 1200 DFBA simulations to train all 24 nsPCE surrogate models. The basis-adaptive hybrid LAR method consistently estimated coefficients in less than 30 seconds, verifying that the DFBA simulations are the dominant cost in this case study. The validation RMSE values are summarized in S2 Table, which are all below the target error threshold.

Fig 8 shows the posterior density estimated using SMC with $N_p = 1 \times 10^6$ particles for a synthetic data set, where the likelihood weights are evaluated using the inexpensive nsPCE surrogate models. The synthetic data ('x' marks in S1 Fig) was obtained by simulating the genome-scale *E. coli* DFBA model with fixed parameters (red lines in Fig 8) and adding random noise realizations (22) to the resulting model outputs. The 1200 DFBA simulations used to construct

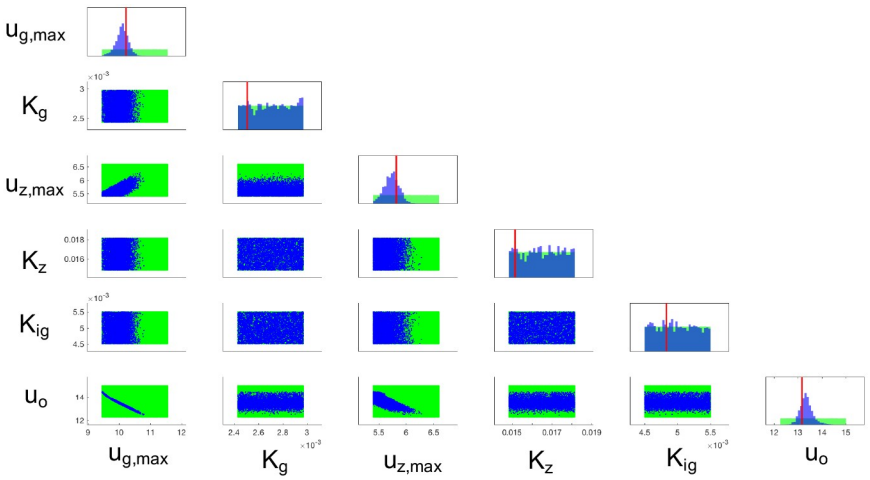

Fig 8. Posterior distribution of the estimated model parameters. The diagonal subplots represent marginal densities while the off-diagonal subplots represent two-dimensional projections of samples from the joint density. Blue denotes the posterior density while green denotes the prior density. The red line represents the true parameter values used to generate synthetic data for estimation purposes.

https://doi.org/10.1371/journal.pcbi.1007308.g008

the surrogates require $\sim$ 30 minutes of CPU time while the surrogate-based SMC algorithm, which takes advantage of vectorization, finishes in $\sim$ 2 minutes of CPU time. Hence, over 800-fold savings in computational cost is achieved when compared to SMC without surrogates that would require approximately 17 days of CPU time under the same settings (1×10^6 DFBA simulations at a cost of 1.5 seconds per evaluation). The DFBA model predictions under the MAP estimates (i.e., parameters that maximize the posterior) are shown in S1 Fig, which closely match the observed data. To verify that the SMC algorithm approximately converged with this many particles, we performed 10 separate bootstrap runs that produced a set of very similar posterior densities. Note that a discussion on challenges and open issues in Bayesian estimation is provided in the Discussion section. The SMC code is provided in the `main_smc.m` script in [29].

The estimated posterior density in Fig 8 provides interesting physical insights. Three of the parameters (K_g, K_z, K_{ig}) are unobservable with the current data set since their posterior (blue) and prior (green) densities are equivalent. This observation could not be easily made before running the estimation procedure due to the nonlinear and indirect relationship between $\boldsymbol{D}$ and $\boldsymbol{X}$. A change in the experimental conditions such as the initial conditions, controlled oxygen concentration, or substrate feed profiles can enhance the sensitivity of the data to parameters (K_g, K_z, K_{ig}). For example, running the batch at low glucose concentrations $g(t) \ll K_g$ results in a glucose uptake rate $u_g(t) \approx u_{g,max} \frac{g(t)}{K_g}$ that is a strong function of K_g, whereas running the batch at high glucose concentrations (as done in this case study) produces a nearly constant uptake rate $u_g(t) \approx u_{g,max}$ that is independent of K_g. Although the data is sensitive to ($u_{g,max}$, $u_{z,max}$, u_o), these parameters are highly correlated as seen in the off-diagonal plots of their joint densities in Fig 8. Thus, the currently available data from one single batch is insufficient for accurately estimating all the parameters of interest. The evolution of the marginal posterior densities of the observable parameters over time is shown in Fig 9. Since glucose is mostly consumed by 7.25 hr, the densities of $u_{g,max}$ and u_o remain constant for the remaining batch time. The density of $u_{z,max}$, however, is constant before 7.25 hr because xylose remains mostly at its initial condition.

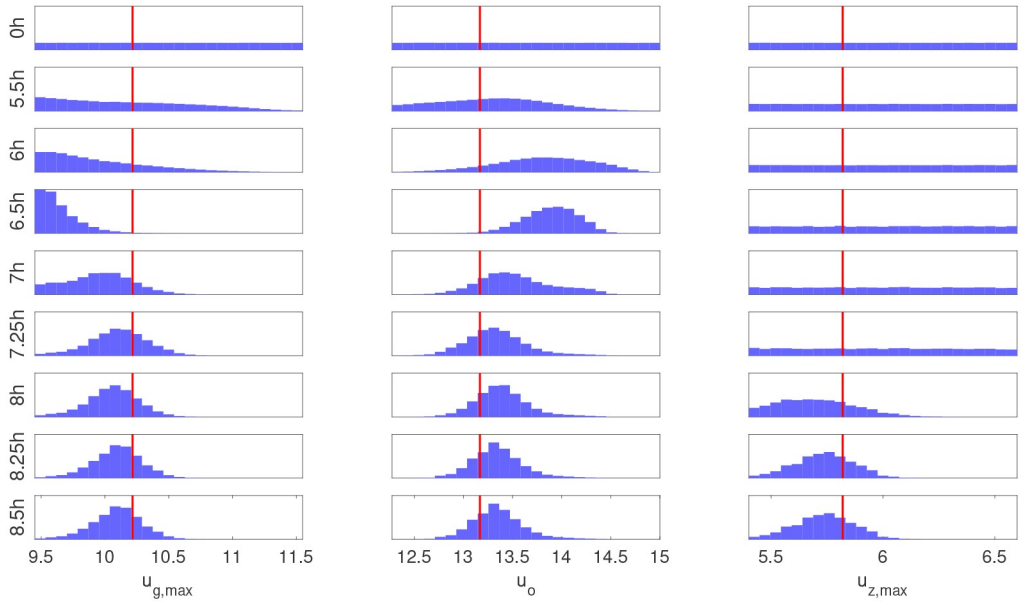

Fig 9. Evolution of the posterior marginal densities for the observable model parameters over time. Each subplot shows the histogram of parameter posterior samples estimated using the sequential Monte Carlo method. The *x*-axis represents the range of values of the parameters and the *y*-axis represents frequencies. The red line represents the true parameter values.

https://doi.org/10.1371/journal.pcbi.1007308.g009

Forward uncertainty propagation. Let $\boldsymbol{Y} = \mathcal{M}(\boldsymbol{X})$ denote the vector of all model responses. The forward UQ problem looks to characterize the uncertainty in the model predictions by propagating uncertainty in the parameters through $\mathcal{M}$. This can involve estimating either the prior predictive distribution $f_{\boldsymbol{Y}}$ (before any data has been collected), or the posterior predictive distribution $f_{\boldsymbol{Y}|\boldsymbol{D}}$ (after data has been obtained). The only difference between these two problems is that $\mathcal{M}$ is evaluated at i.i.d. samples drawn from the prior in the former and the posterior in the latter. The densities of the model predictions estimated using 1×10^6 samples are shown in Fig 10. By replacing the full DFBA model with the nsPCE surrogate model, these histograms were obtained in less than 1 minute of CPU time. As expected, the prior predictive distributions are much wider than the posterior predictive distributions, indicating there is significant uncertainty in the predictions before incorporating data. In addition, we see that many of these distributions have sharp changes and long tails due to the non-smooth behavior of the model responses, which can be accurately captured with the proposed nsPCE framework. It is also interesting to note that the posterior predictive distributions have low variance, even though the parameters are not perfectly estimated. This highlights the impact that nonlinearity can have on both estimation and uncertainty propagation.

Case study 2: Synthetic metabolic network

This case study is based on a synthetic metabolic network originally introduced in [31, Chapter 8]. The goal of this case study is to show that the proposed nsPCE method can be applied to problems with a larger number of parameters as well as alternative UQ approaches. The synthetic metabolic network consumes a carbon source *C*, a nitrogen

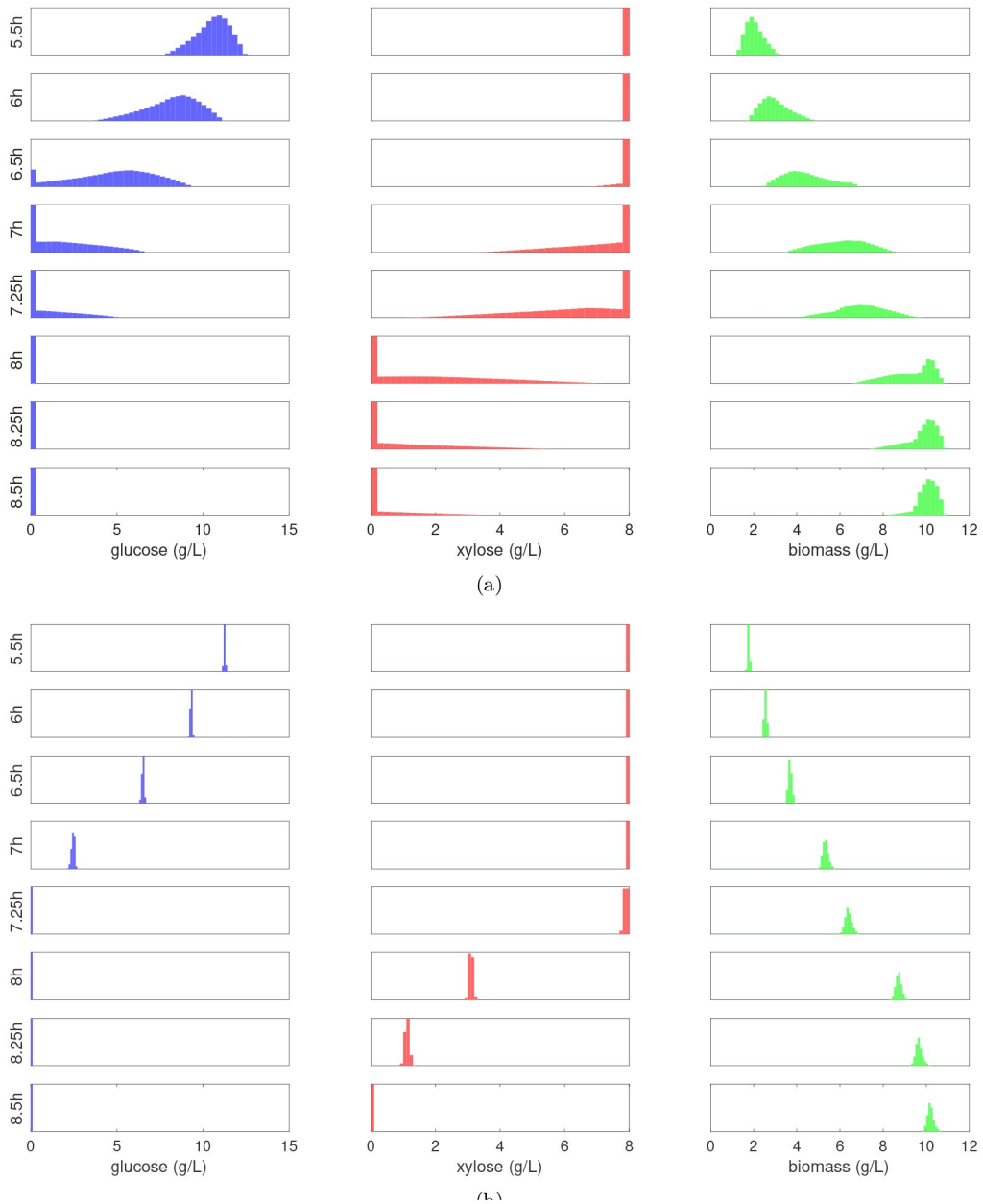

Fig 10. Predicted probability distributions of extracellular concentrations. (a) Model predictions using parameter samples from the prior. **(b)** Model predictions using parameter samples from the posterior. Each subplot shows the histogram of samples of the model output obtained by substituting i.i.d. samples from the parameter distribution into the corresponding ME-PCE surrogate model. The *x*-axis represents the range of values of the model outputs and the *y*-axis represents frequencies.

https://doi.org/10.1371/journal.pcbi.1007308.g010

source N, and an oxygen source O to produce lipids L, ethanol E, biomass X, ATP, and some oxidation product COX. Although used for illustrative purposes, this network is meant to mimic the behavior of living organisms in the sense that: (i) E can only be produced in the absence of O, (ii) L can only be accumulated in the absence of N, (iii) there is a minimum ATP requirement, and (iv) the aerobic oxidation of C produces more energy than

fermentation of C. The set of reactions can be summarized as

$$\begin{aligned}
v_C &: C_{ex} \rightarrow C,\\
v_N &: N_{ex} \rightarrow N,\\
v_O &: O_{ex} \rightarrow O,\\
v_{OX} &: C + O \rightarrow S_{ATP,OX}ATP + S_{OX,OX}COX_{ex},\\
v_{Ferm} &: 4C \rightarrow S_{ATP,Ferm}ATP + E_{ex} + S_{OX,Ferm}COX_{ex},\\
v_L &: 4C + S_{ATP,L}ATP \rightarrow L,\\
v_X &: S_{C,X}C + S_{N,X}N + S_{ATP,X}ATP \rightarrow X,\\
v_{ATP} &: ATP \rightarrow ATP_{maintenance},
\end{aligned} \tag{25}$$

where the subscript ex denotes extracellular metabolites and all of the reactions are assumed to be unidirectional. The unknown stoichiometric coefficients are denoted by $S_{i,j}$, where i represents the metabolite name and j represents the reaction name. The dynamic mass balance equations for the extracellular environment are given by

$$\begin{aligned}
\dot{X}(t) &= v_X(t)X(t), & X(0) &= X_0,\\
\dot{C}(t) &= -v_C(t)X(t), & C(0) &= C_0,\\
\dot{N}(t) &= -v_N(t)X(t), & N(0) &= N_0,\\
\dot{O}(t) &= -v_O(t)X(t), & O(0) &= O_0,\\
\dot{L}(t) &= v_L(t)X(t), & L(0) &= 0,\\
\dot{E}(t) &= v_{Ferm}(t)X(t), & E(0) &= 0,\\
\dot{COX}(t) &= (S_{OX,OX}v_{OX}(t) + S_{OX,Ferm}v_{Ferm}(t))X(t), & COX(0) &= 0,\\
\dot{\alpha}(t) &= \gamma(t), & \alpha(0) &= 0,
\end{aligned} \tag{26}$$

where α is a penalty state that remains equal to zero until the state trajectories become infeasible (e.g., when all of the metabolites are depleted). A detailed discussion on how to determine the instantaneous penalty value γ is provided in [10], which is automatically computed in `DFBAlab`. We assume that the uptake kinetics are given by the following expressions

$$\begin{aligned}
v_C^{UB}(\mathbf{s}) &= \max\left(0, v_{max,C}\frac{C}{K_C + C}\frac{1}{1 + \frac{E}{K_{iE}}}\right),\\
v_N^{UB}(\mathbf{s}) &= \max\left(0, v_{max,N}\frac{N}{K_N + N}\right),\\
v_O^{UB}(\mathbf{s}) &= \max\left(0, v_{max,O}\frac{O}{K_O + O}\right),
\end{aligned} \tag{27}$$

where $\mathbf{s} = (X, C, N, O, L, E, COX, \alpha)$ is the vector of extracellular species. A hierarchical set of objectives is used in the FBA problem (2) to ensure that unique reaction fluxes are obtained (see S3 Table). A total of twenty parameters in this DFBA model, appearing in both

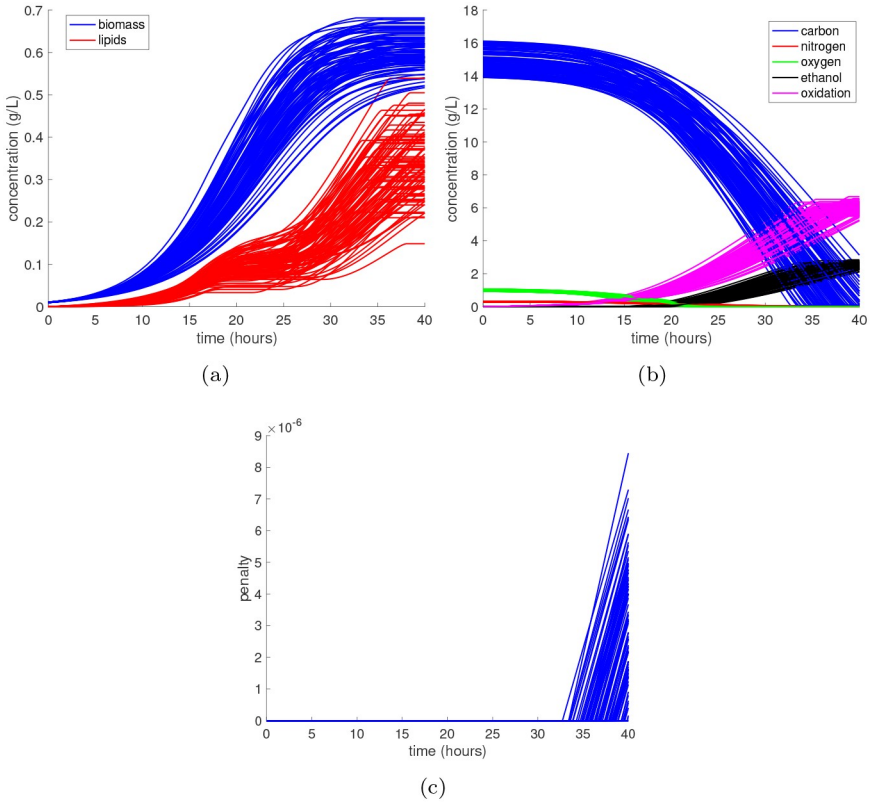

Fig 11. Monte Carlo simulation of the synthetic metabolic network. The synthetic DFBA model with twenty uncertain parameters is integrated from time 0 to 40 hours for 100 different parameter realizations drawn independently from the uniform prior density. The time profiles are shown for **(a)** biomass and lipids, **(b)** the substrates and products, and **(c)** the penalty state.

https://doi.org/10.1371/journal.pcbi.1007308.g011

intracellular and extracellular quantities, are assumed to be uniformly distributed between upper and lower bounds summarized in S4 Table.

Global sensitivity analysis. To locate any possible singularities, we first simulate the DFBA model with randomly sampled parameter values. The results are shown in Fig 11 wherein we see that the penalty state becomes positive $\alpha(t) > 0$ once all substrates are depleted, which introduces a strong discontinuity into the state profiles. Even though this is a considerably smaller metabolic network than the one considered in the *E. coli* case study, it still takes approximately 0.5 seconds of CPU time per realization of the parameter. Thus, it is still advantageous to construct a surrogate model to speedup both forward and inverse UQ problems.

We look to run the proposed nsPCE method (see Fig 1) using the time that the penalty state switches from zero to positive as the singularity time. As suggested in [31, Chapter 8], we consider the seven substrate and product concentrations (X, C, N, O, L, E, COX) at four time points $t \in \{10, 20, 30, 40\}$ hr as our main quantities of interest. The nsPCE method was applied in the same manner as described in the previous case study. Here, we specified a target error of $\varepsilon_{target} = 10^{-3}$, 100 initial ED samples, 100 ED samples added at each iteration, 1500 maximum ED samples, the maximum polynomial degree could vary from 1 to 30, and a hyperbolic truncation scheme with $q = 0.6$. The algorithm converged using a total of 2800 DFBA simulations. The resulting parity plots are shown in S2 Fig, which all have empirical RMSE values significantly below the target error. To further assess the accuracy of these models, we calculated the RMSE using 1000 additional samples that were not used during the training process. The validation

RMSE averaged over the 28 models was found to be 8.1×10^{-4}. Only 6 of the 28 surrogates had RMSE values slightly above the target, with the largest overall RMSE being 3.5×10^{-3}, indicating that the surrogates are reasonably accurate representations of the original model. Note that these errors can be refined by specifying a lower ε_{target} at the cost of more DFBA simulations.

Once the nsPCE surrogate models are constructed, they can be used to efficiently perform global sensitivity analysis in order to quantify the respective effects of each individual parameter on the variance of the model response. Although many sensitivity measures exist, we use Sobol' indices since they make no assumption on the underlying linearity or monotonicity of the model. The global sensitivity results for the various quantities of interest over time are shown in Fig 12. A variety of interesting conclusions can be drawn from these results. For example, the model appears to be insensitive to K_C and K_{iE}, which is likely due to the fact that the batch was run at high carbon and low ethanol concentrations. In addition, the measurements of carbon, nitrogen, and oxygen are highly sensitive to their respective initial conditions C_0, N_0, and O_0 at the first measurement time of 10 hr; however, this sensitivity drops considerably as time evolves. We emphasize that obtaining such insights using random sampling on the full model can be prohibitively expensive, but requires negligible cost using the nsPCE surrogate models.

Optimization-based parameter estimation. We now utilize maximum a posteriori (MAP) estimation to estimate the unknown model parameters from synthetically-generated experimental data (see S5 Table). The MAP estimate is defined as the mode of the posterior distribution, and can be stated directly as an optimization problem of the form [53]

$$\hat{\boldsymbol{x}}_{\mathrm{MAP}}(\boldsymbol{d}) = \underset{\boldsymbol{x}\in S}{\operatorname{argmax}}\, f_{X|D}(\boldsymbol{x}|\boldsymbol{d}) = \underset{\boldsymbol{x}\in S}{\operatorname{argmax}}\, f_{D|X}(\boldsymbol{d}|\boldsymbol{x}) f_X(\boldsymbol{x}), \tag{28}$$

where the prior acts as a regularization term that can stabilize the solution whenever the parameters cannot be uniquely inferred from the available data [54]. We consider a Gaussian likelihood, with noise standard deviations reported in S5 Table, and a Gaussian prior whose mean is equal to the midpoint of the bounds in S4 Table and standard deviations equal to 10% of the mean values. Under the Gaussian restrictions, we can convert the MAP problem to the minimization of a regularized weighted least squares objective by applying a negative log transformation. We solved the optimization (28) using both the full DFBA and nsPCE surrogate models in order to assess the computational gains afforded by the nsPCE method. To ensure a fair comparison, we solved both of these MAP problems in Matlab using the non-smooth optimizer SolvOpt [55] with default parameter settings and the mean of the prior as the initial guess. The algorithm took approximately 2.5 hr to converge when using the full DFBA model, which was substantially reduced to less than 2 minutes (i.e., a factor of 60) when the full model was replaced with the nsPCE surrogates.

Not only did the use of the nsPCE surrogate models accelerate the optimization, it also produced a solution with a lower overall objective function value. The objective improved from 471.73 to 1.56 when using the surrogate models as compared to 63.57 when using the full DFBA model. Convergence to a suboptimal solution is likely a consequence of numerical issues related to the stability of derivative approximation using finite difference in DFBA models, which were also observed in [31, Chapter 8]. On the other hand, since the nsPCE surrogate models are defined in terms of simple polynomial functions, the finite difference derivative approximation seems to produce more stable iterations towards the solution of the MAP problem, at least in this particular case study. The predictions of the DFBA model under the MAP estimates found using the nsPCE surrogates are shown in Fig 13. We see that the predictions using the posterior parameter estimates very closely match the observed data, which is a large improvement when compared to the predictions based on the prior parameter estimates.

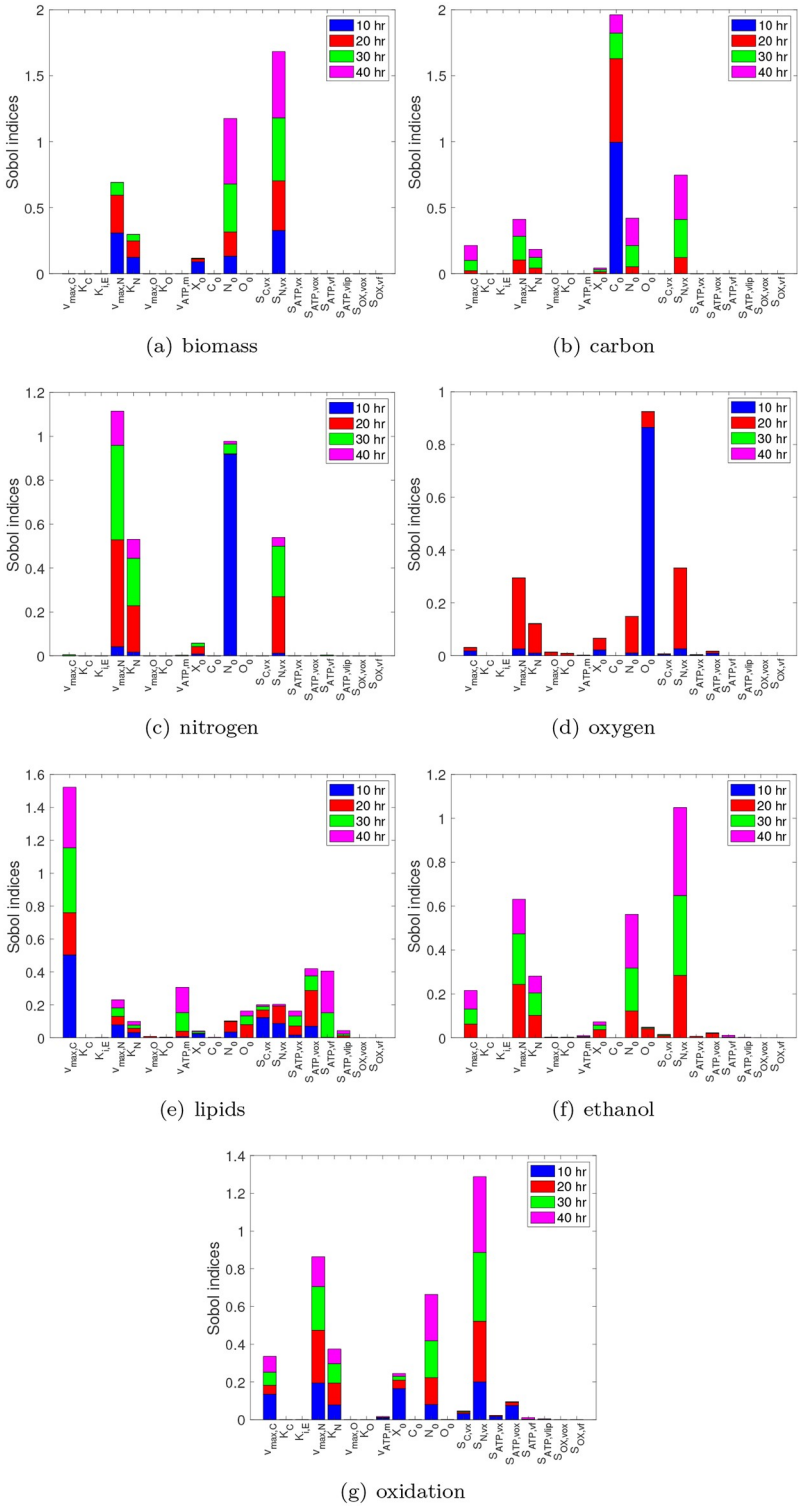

Fig 12. Global sensitivity indices for the quantities of interest in the synthetic metabolic network. (a)-(g) Global sensitivity indices for extracellular substrate and product concentrations at various time points with respect to the twenty uncertain parameters.

https://doi.org/10.1371/journal.pcbi.1007308.g012

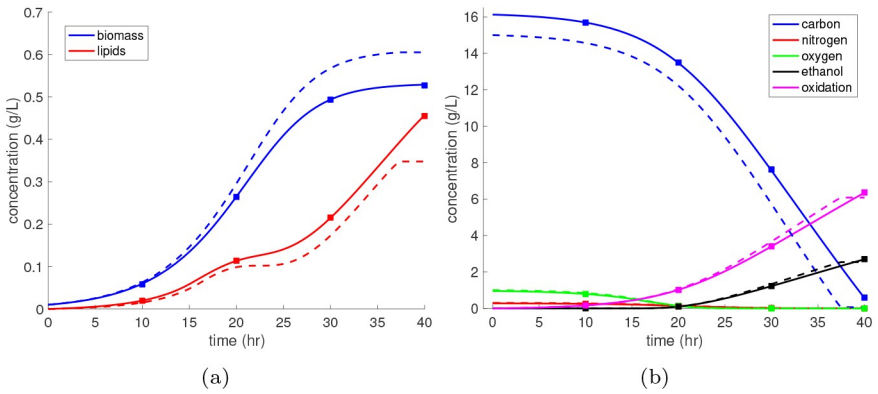

Fig 13. Comparison of model predictions and data for the synthetic metabolic network. The model predictions for **(a)** biomass and lipids and **(b)** the substrates and products, shown with solid lines, were obtained by integrating the DFBA model with the MAP estimates of the parameters. The '□' marks represent the data that was obtained by corrupting the model predictions using the true (unknown) parameters with randomly generated noise. The dotted lines represent the model predictions based on the initial parameter guess that was used to initialization the optimizer.

https://doi.org/10.1371/journal.pcbi.1007308.g013

Posterior distribution analysis. The MAP estimation (28) determines the parameters that maximize the posterior density. However, we need a characterization of the entire posterior to assess uncertainty in these estimates. Here, we use a Laplace approximation of the posterior density, which is based on a second-order Taylor series of $-\log(f_{X|D}(\boldsymbol{x}|\boldsymbol{d}))$ around the MAP estimate [56]. As shown in [57], this leads to a Gaussian approximation of the posterior whose mean is equal to the MAP estimate and whose covariance is defined in terms of the model response sensitives. The approximated posterior marginal densities and 95% confidence regions for the twenty MAP parameter estimates are shown in Fig 14. As can be seen, the true (unknown) parameter values are contained within the reported confidence regions. We also see that the parameters with the highest global sensitivity indices (see Fig 12) are accurately

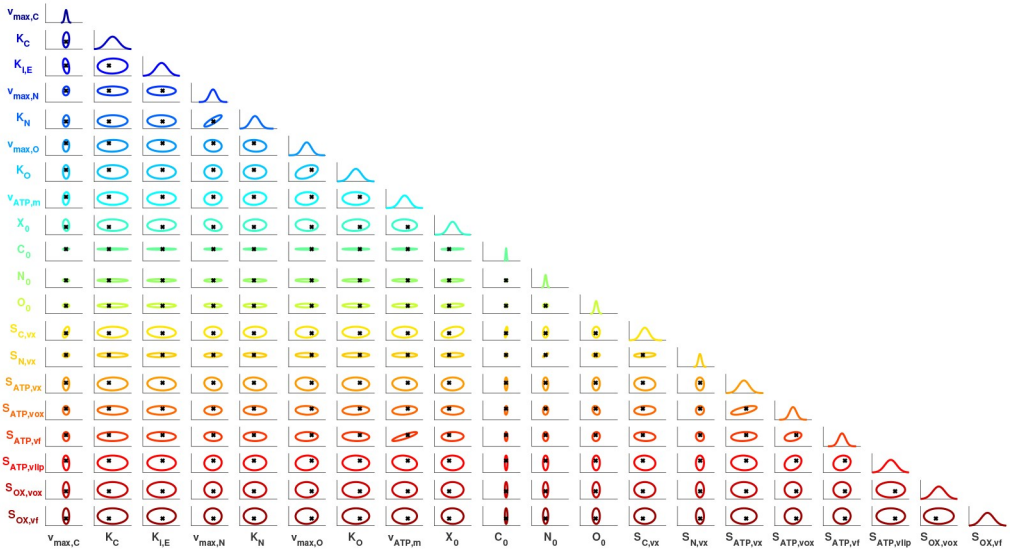

Fig 14. Estimated posterior distribution for the parameters of the synthetic metabolic network. The diagonal subplots represent the estimated marginal densities, while the off-diagonal subplots represent the two-dimensional projections of the 95% confidence regions. Black 'x' marks represent the true parameter values, while the modes of the marginal densities signify the MAP estimates.

https://doi.org/10.1371/journal.pcbi.1007308.g014

estimated, whereas the parameters that have little-to-no sensitivity to the data have much wider variances that are similar to that of the prior. Lastly, we observe that physically-related parameters exhibit a significant degree of correlation including, for example, nitrogen uptake parameters $v_{max,N}$ and K_N. It is worth noting that the surrogate models can also enable the use of more advanced methods for posterior characterization such as randomized MAP [58], which would require the repeated solution of (28) with randomly perturbed data.

Discussion

In this work, we develop a novel surrogate modeling method for handling the non-smooth nature of computationally expensive dynamic flux balance analysis (DFBA) models. It is shown that surrogate models can vastly accelerate uncertainty quantification (UQ) tasks, such as calibrating the model with experimental data (inverse problem) and quantifying confidence in the model predictions (forward problem). The proposed surrogate modeling method is based on an extension of polynomial chaos expansion (PCE), which we refer to as non-smooth PCE (nsPCE). The main idea behind nsPCE is to systematically partition the parameter space into two non-overlapping regions (or elements) on which the model response behaves smoothly. The nsPCE uses a model of the time that the singularity occurs in order to define the boundary between these two elements. State-of-the-art (i.e., sparse basis-adaptive) regression methods are used to estimate the coefficients of the expansions, such that the overall model response is approximated by a sparse piecewise polynomial function.

We demonstrate the advantages of the nsPCE surrogate modeling method on two separate case studies. The first case study is based on a DFBA model of an *E. coli* fermentation reactor under aerobic growth in a glucose and xylose mixed media. A genome-scale metabolic network reconstruction with 1075 reactions and 761 metabolites is used to represent the intracellular behavior, which results in an expensive-to-simulate DFBA model that is prohibitive for use in most UQ tasks. Thus, we illustrate how both inverse and forward UQ can be significantly accelerated using nsPCE surrogate models on this problem. In particular, we use a Bayesian estimation method to infer six uncertain parameters related to the substrate uptake kinetics from data. The posterior parameter distribution is estimated using sequential Monte Carlo with 1×10^6 samples, which would have required ~ 17 days of CPU time to compute using the full DFBA model, but takes less than one hour when using the nsPCE surrogate models including the cost of training the models. The resulting posterior distribution yields significant physical insights including that the available data set is insufficient to reliably estimate all six parameters, with three of the parameters being non-identifiable under the current experimental conditions. We then demonstrate the scalability of the proposed nsPCE method on a synthetic metabolic network problem with twenty unknown parameters that are related to both intracellular and extracellular quantities. We estimate these parameters using maximum a posteriori (MAP) estimation, and observe that the cost of the optimization algorithm can be reduced by a factor of 60 when using the nsPCE surrogates in place of the full DFBA model. Note that the observed speedups are expected to be even greater for more complex DFBA models, such as those with nonlinear cellular objectives, multiple cultures, or even larger metabolic networks due to the increased cost of the simulations.

Scalability properties of nsPCE

The nsPCE method is specifically constructed to take advantage of the hybrid LAR method for sparse regression, which was originally developed in [19]. As such, nsPCE directly inherits the beneficial scalability properties of hybrid LAR that introduces two sources of sparsity into the expansions: (i) low-rank truncation that discards basis terms that lead to high-order

interaction of the parameters that are irrelevant in most engineering problems and (ii) regularized least squares is used to systematically add basis terms that are strongly correlated to the model response. Additionally, the risk of over-fitting the surrogate model to the available data set can be reduced even further by making the approach basis-adaptive, i.e., separate PCE models are fit for varying maximum degrees and the one with the lowest error is selected.

The basis-adaptive hybrid LAR approach has been successfully applied to a wide-variety of problems and has consistently shown the ability to greatly mitigate the curse-of-dimensionality that is inherent in traditional PCE methods (see, for example, [19, 59, 60]). To the best of our knowledge, [60] tackled the largest problem to-date, which is a hydrogeological model with 78 parameters (68 identified to be sensitive) that can be accurately represented using a sparse PCE trained using only 2000 model evaluations. Although uncertainty in high-dimensional DFBA models has not been explored in the literature, these promising results and those shown in the synthetic case study give some confidence that nsPCE may be able to scale to the sizes needed to solve these challenging problems. Note that very recent work has shown that sparse PCEs can be applied to ultrahigh-dimensional problems (on the order of 10^4 parameters) by incorporating a dimensionality reduction step before training the surrogate model [61]. It may be possible to use similar approaches to incorporate uncertainty in the complete set of intracellular model parameters into the nsPCE surrogate models. These are interesting and important challenges that deserve further investigation.

Further reducing the number of model evaluations

In this work, the surrogate models are trained using experimental designs (EDs) populated with random samples of the parameters. Recent work has demonstrated that the number of ED points needed to achieve a desired accuracy level can be further reduced by maximizing the information content of the sample locations. Multiple approaches have been developed to tackle this challenging problem, including coherence-optimal sampling [62] and numerical "moment-matching" optimization [34, 37]. The optimal placement of samples in arbitrary domain shapes in a sequential fashion remains largely unexplored in the literature.

Additionally, the current implementation of nsPCE involves only two elements; however, it is unclear if the convergence rate can be improved even more by further decomposing these elements. An adaptive approach for decomposing the random parameter space that uses sensitivity information to decide which elements to split was proposed in [25]. A similar concept could be potentially utilized within nsPCE, though the method would likely benefit from the incorporation of more advanced geometries than simple boxes.

Considerations and challenges in parameter estimation

Many of the difficulties encountered during parameter estimation are related to poor identifiability of model parameters. Performing parameter identifiability tests can help mitigate these difficulties by ensuring the parameter estimation problem is well-posed, which is especially important when dealing with limited experimental data and/or considering a large number of model parameters. It is common to distinguish between structural and practical identifiability. Structural identifiability is a theoretical property of the model structure that depends only on the observation function and the manipulated input function. Since a structurally non-identifiable parameter is independent of the accuracy of available experimental data, it cannot be resolved by a refinement of existing measurements. The only remedy is a qualitatively new measurement or experiment that alters the structure of the mapping between the parameters and the data. In contrast, practical identifiability also takes into account the amount and quality of the measured data, meaning that it can in principle be

resolved by improving the quality of the measurements or increasing the number of measured time points. A thorough treatment of these issues in the context of biological models can be found in, e.g., [63–65]. To the best of our knowledge, structural and practical identifiability analysis has not been demonstrated on DFBA models, which is an interesting area for future work. It is important to note that, although many methods exist for detecting non-identifiable parameters, they often have restrictions on the class of functions so that they are not directly applicable to DFBA models.

Although not observed here, sequential Monte Carlo (SMC) can suffer from *degeneracy* wherein fewer and fewer particles retain significant weight. This is especially prevalent in high-dimensional problems including those with a large number of parameters or a large time horizon [66]. In [67], it is shown that the degeneracy phenomenon occurs unless the sample size is chosen to be exponential in the dimension, which indicates some type of curse-of-dimensionality. This sample degeneracy can be protected against by adding a *rejuvenation* step that "moves" the resampled particles according to a Markov chain Monte Carlo (MCMC) transition kernel [51]. This operation does not change the target distribution, but does reduce impoverishment since identical replicates of a single particle are replaced with new values. The most challenging part of the MCMC step is ensuring that the samples obtained realistically represent the desired distribution. It is known that convergence of the Markov chain fails for posteriors that are not proper, which can happen whenever the prior is improper (e.g., uniform density with infinite bounds) or non-identifiable parameters exist in the model [68]. In these situations, neither the prior assumptions nor the likelihood that represents the experimental data sufficiently constrain the posterior distribution. As such, the convergence properties of SMC and MCMC methods may improve considerably by resolving parameter identifiability issues before running the algorithm [69].

Extensions to optimal experiment design

The selection of optimal conditions for conducting experiments (e.g., measurement times, initial conditions, and time-varying input profiles) is important for ensuring maximum information is extracted from the observations, especially when the experiments are expensive and time-consuming to perform. For example, it may be useful to change the feed rate or the measurement times in the considered case study so that the data ensures tight parameter estimates are obtained. Optimal experiment design (OED) has been extensively studied in the classical framework wherein the design criteria are defined as some scalar function of the Fisher information matrix (FIM) [70, 71]. More recently, OED has been tackled from a fully Bayesian perspective that replaces the approximated classical design criteria with an *expected utility* function that is rigorously chosen from a decision-theoretic point of view [72–74].

The nsPCE surrogate models could be used to efficiently evaluate classical or Bayesian design criteria at any fixed experimental condition. However, the parameter space decomposition depends strongly on the experiment, such that separate surrogates need to be constructed for all experiments of interest. This is not a major challenge when only a small number of experiments are considered, but may become intractable for continuous design spaces. Developing efficient procedures for both classical and Bayesian OED in genome-scale DFBA models is an important area for future research. One possible direction is to treat the experiment design variables as parameters when constructing the surrogate model, as suggested in [75] for global PCE. It would be interesting to see how well nsPCE can handle these additional dimensions, since the model responses would likely be highly sensitive to the design variables.

Supporting information

S1 Fig. Comparison of model predictions and synthetic data for the *E. coli* case study. The model predictions, shown with solid lines, were obtained by integrating the DFBA model with the maximum a posteriori (MAP) estimates of the parameters, which correspond to the mode of the posterior density. The 'x' marks represent synthetic data generated by corrupting model predictions for the true (unknown) parameters with randomly generated noise.
(PDF)

S2 Fig. Parity plots for nsPCE surrogate models for synthetic metabolic network. The rows correspond to the extracellular substrate and product concentrations while the columns correspond to the various time points of interest. The x-axis represents the exact value of the model while the y-axis represents the surrogate model predictions. The blank plots represent quantities of interest with variance significantly lower than the tolerance.
(PDF)

S1 Table. Nominal substrate uptake parameters for *E. coli* DFBA model. Parameter values taken from [8]. Uncertainty in the parameter estimates was not quantified. We assume the uncertainty in these estimates is uniformly distributed around ±10% of the nominal parameter values, which leads to fairly large variability in the predicted extracellular behavior.
(PDF)

S2 Table. Relative mean square error estimates for nsPCE surrogate models for *E. coli* case study. A total of 1200 DFBA simulations were sequentially generated to train all nsPCE surrogate models to meet the specification $\varepsilon_{target} = 10^{-3}$. The RMSE values were computed using a validation set of 10,000 DFBA simulations. An entry of 0.0 corresponds to quantities of interest with variance below ε_{target}.
(PDF)

S3 Table. Hierarchy of objectives for synthetic metabolic network. The FBA problem was formulated as a linear program with multiple objectives that are optimized based on the priority list specified in this table. This approach is able to ensure that the FBA problem is feasible for all simulation times and that the exchange fluxes are unique. More information on this strategy can be found in [10].
(PDF)

S4 Table. Uncertain parameter bounds for synthetic metabolic network. The uncertainty ranges are based on the nominal values presented in [31, Chapter 8]. The parameters are either related to the substrate uptake kinetics, the initial conditions, or the stoichiometric coefficients of the reactions. For the latter, we selected coefficients that are likely to be inferred from experimental data in real applications.
(PDF)

S5 Table. Simulated experimental data for synthetic metabolic network. This data was used to estimate the parameters in the synthetic case study, which was obtained by simulating the DFBA model with true (unknown) parameters and then adding randomly generated noise. The noise was assumed to be Gaussian with standard deviation shown in row labeled 'STDEV'.
(PDF)

S1 Text. Summary of methods for simulating DFBA models.
(PDF)

Author Contributions

Writing – original draft: Joel A. Paulson, Marc Martin-Casas, Ali Mesbah.

References

1. Covert MW, Knight EM, Reed JL, Herrgard MJ, Palsson BO. Integrating high-throughput and computational data elucidates bacterial networks. Nature. 2004; 429(6987):92–96. https://doi.org/10.1038/nature02456 PMID: 15129285
2. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? Nature Biotechnology. 2010; 28:245–248. https://doi.org/10.1038/nbt.1614 PMID: 20212490
3. O'Brien EJ, Monk JM, Palsson BO. Using genome-scale models to predict biological capabilities. Cell. 2015; 161(5):971–987. https://doi.org/10.1016/j.cell.2015.05.019 PMID: 26000478
4. Van Gulik WM, Heijnen JJ. A metabolic network stoichiometry analysis of microbial growth and product formation. Biotechnology and Bioengineering. 1995; 48:681–698. https://doi.org/10.1002/bit.260480617 PMID: 18623538
5. Mahadevan R, Edwards JS, Doyle FJ III. Dynamic flux balance analysis of diauxic growth in Escherichia coli. Biophysical Journal. 2002; 83:1331–1340. https://doi.org/10.1016/S0006-3495(02)73903-9 PMID: 12202358
6. Hjersted JL, Henson MA. Optimization of fed-batch Saccharomyces cerevisiae fermentation using dynamic flux balance models. Biotechnology Progress. 2006; 22:1239–1248. https://doi.org/10.1021/bp060059v PMID: 17022660
7. Meadows AL, Karnik R, Lam H, Forestell S, Snedecor B. Application of dynamic flux balance analysis to an industrial Escherichia coli fermentation. Metabolic engineering. 2010; 12:150–160. https://doi.org/10.1016/j.ymben.2009.07.006 PMID: 19646545
8. Hanly TJ, Henson MA. Dynamic flux balance modeling of microbial co-cultures for efficient batch fermentation of glucose and xylose mixtures. Biotechnology and Bioengineering. 2011; 108:376–385. https://doi.org/10.1002/bit.22954 PMID: 20882517
9. Lisha KP, Sarkar D. Dynamic flux balance analysis of batch fermentation: Effect of genetic manipulations on ethanol production. Bioprocess and Biosystems Engineering. 2014; 37:617–627. https://doi.org/10.1007/s00449-013-1027-y PMID: 23921448
10. Gomez JA, Höffner K, Barton PI. DFBAlab: A fast and reliable MATLAB code for dynamic flux balance analysis. BMC Bioinformatics. 2014; 15:409. https://doi.org/10.1186/s12859-014-0409-8 PMID: 25519981
11. Kennedy MC, O'Hagan A. Bayesian calibration of computer models. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2001; 63:425–464. https://doi.org/10.1111/1467-9868.00294
12. Ghanem R, Spanos P. Stochastic Finite Elements A Spectral Approach. SpringerVerlag; 1991.
13. Xiu D. Fast numerical methods for stochastic computations: A review. Communications in Computational Physics. 2009; 5:242–272.
14. Smith RC. Uncertainty quantification: Theory, implementation, and applications. vol. 12. SIAM; 2013.
15. Marzouk YM, Najm HN. Dimensionality reduction and polynomial chaos acceleration of Bayesian inference in inverse problems. Journal of Computational Physics. 2009; 228:1862–1902. https://doi.org/10.1016/j.jcp.2008.11.024
16. Scott F, Wilson P, Conejeros R, Vassiliadis VS. Simulation and optimization of dynamic flux balance analysis models using an interior point method reformulation. Computers & Chemical Engineering. 2018; 119:152–170. https://doi.org/10.1016/j.compchemeng.2018.08.041
17. Xiu D, Karniadakis GE. The Wiener–Askey polynomial chaos for stochastic differential equations. SIAM Journal of Scientific Computing. 2002; 24:619–644. https://doi.org/10.1137/S1064827501387826
18. Xiu D. Efficient collocational approach for parametric uncertainty analysis. Communications in Computational Physics. 2007; 2:293–309.
19. Blatman G, Sudret B. Adaptive sparse polynomial chaos expansion based on least angle regression. Journal of Computational Physics. 2011; 230:2345–2367. https://doi.org/10.1016/j.jcp.2010.12.021
20. Paulson JA, Mesbah A. An efficient method for stochastic optimal control with joint chance constraints for nonlinear systems. International Journal of Robust and Nonlinear Control. 2017; p. 1–21.
21. Streif S, Petzke F, Mesbah A, Findeisen R, Braatz RD. Optimal experimental design for probabilistic model discrimination using polynomial chaos. IFAC Proceedings Volumes. 2014; 47(3):4103–4109. https://doi.org/10.3182/20140824-6-ZA-1003.01562

22. Martin-Casas M, Mesbah A. Discrimination between competing model structures of biological systems in the presence of population heterogeneity. IEEE Life Science Letters. 2016; 2:23–26. https://doi.org/10.1109/LLS.2016.2644645

23. Renardy M, Yi TM, Xiu D, Chou CS. Parameter uncertainty quantification using surrogate models applied to a spatial model of yeast mating polarization. PLoS Computational Biology. 2018; 14: e1006181. https://doi.org/10.1371/journal.pcbi.1006181 PMID: 29813055

24. Wan X, Karniadakis GE. An adaptive multi-element generalized polynomial chaos method for stochastic differential equations. Journal of Computational Physics. 2005; 209:617–642. https://doi.org/10.1016/j.jcp.2005.03.023

25. Wan X, Karniadakis GE. Multi-element generalized polynomial chaos for arbitrary probability measures. SIAM Journal on Scientific Computing. 2006; 28:901–928. https://doi.org/10.1137/050627630

26. Höffner K, Harwood SM, Barton PI. A reliable simulator for dynamic flux balance analysis. Biotechnology and Bioengineering. 2013; 110:792–802. https://doi.org/10.1002/bit.24748 PMID: 23055276

27. Harwood SM, Höffner K, Barton PI. Efficient solution of ordinary differential equations with a parametric lexicographic linear program embedded. Numerische Mathematik. 2016; 133:623–653. https://doi.org/10.1007/s00211-015-0760-3

28. Reed JL, Vo TD, Schilling CH, Palsson BO. An expanded genome-scale model of Escherichia coli K-12 (i JR904 GSM/GPR). Genome Biology. 2003; 4:R54. https://doi.org/10.1186/gb-2003-4-9-r54 PMID: 12952533

29. Paulson, JA. The nsPCE toolbox. https://github.com/joelpaulson/nsPCE.

30. Zhao X, Noack S, Wiechert W, von Lieres E. Dynamic flux balance analysis with nonlinear objective function. Journal of Mathematical Biology. 2017; 75:1487–1515. https://doi.org/10.1007/s00285-017-1127-4 PMID: 28401266

31. Gomez JA. Simulation, sensitivity analysis, and optimization of bioprocesses using dynamic flux balance analysis. Massachusetts Institute of Technology; 2018.

32. Soize C, Ghanem R. Physical systems with random uncertainties: Chaos representations with arbitrary probability measure. SIAM Journal on Scientific Computing. 2004; 26:395–410. https://doi.org/10.1137/S1064827503424505

33. Gautschi W. On generating orthogonal polynomials. SIAM Journal on Scientific and Statistical Computing. 1982; 3(3):289–317. https://doi.org/10.1137/0903018

34. Paulson JA, Mesbah A. Arbitrary Polynomial Chaos for Quantification of General Probabilistic Uncertainties: Shaping Closed-loop Behavior of Nonlinear Systems. In: Proceedings of the 57th IEEE Conference on Decision and Control. Miami; 2018. p. Accepted.

35. Rosenblatt M. Remarks on a multivariate transformation. The Annals of Mathematical Statistics. 1952; 23:470–472. https://doi.org/10.1214/aoms/1177729394

36. Paulson JA, Buehler EA, Mesbah A. Arbitrary Polynomial Chaos for Uncertainty Propagation of Correlated Random Variables in Dynamic Systems. IFAC-PapersOnLine. 2017; 50:3548–3553. https://doi.org/10.1016/j.ifacol.2017.08.954

37. Paulson JA, Mesbah A. Nonlinear Model Predictive Control with Explicit Backoffs for Stochastic Systems under Arbitrary Uncertainty. In: Proceedings of the 6th IFAC Conference on Nonlinear Model Predictive Control. Madison, WI; 2018. p. 622–633.

38. Feinberg J, Eck VG, Langtangen HP. Multivariate polynomial chaos expansions with dependent variables. SIAM Journal on Scientific Computing. 2018; 40(1):A199–A223. https://doi.org/10.1137/15M1020447

39. Constantine PG, Eldred MS, Phipps ET. Sparse pseudospectral approximation method. Computer Methods in Applied Mechanics and Engineering. 2012; 229:1–12. https://doi.org/10.1016/j.cma.2012.03.019

40. Sinsbeck M, Nowak W. An optimal sampling rule for nonintrusive polynomial chaos expansions of expensive models. International Journal for Uncertainty Quantification. 2015; 5:275–295. https://doi.org/10.1615/Int.J.UncertaintyQuantification.2015008446

41. Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression. The Annals of Statistics. 2004; 32:407–499. https://doi.org/10.1214/009053604000000067

42. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics. 2005; 21:3301–3307. https://doi.org/10.1093/bioinformatics/bti499 PMID: 15905277

43. Marelli S, Sudret B. UQLab: A framework for uncertainty quantification in Matlab. In: Vulnerability, Uncertainty, and Risk: Quantification, Mitigation, and Management; 2014. p. 2554–2563.

44. Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, Herrgard MJ. Quantitative prediction of cellular metabolism with constraint-based models: The COBRA toolbox. Nature Protocols. 2007; 2:727. https://doi.org/10.1038/nprot.2007.99 PMID: 17406635

45. Mao L, Verwoerd WS. ORCA: a COBRA toolbox extension for model-driven discovery and analysis. Bioinformatics. 2013; 30:584–585. https://doi.org/10.1093/bioinformatics/btt723 PMID: 24336807

46. Suykens JAK, Vandewalle J. Least squares support vector machine classifiers. Neural Processing Letters. 1999; 9:293–300. https://doi.org/10.1023/A:1018628609742

47. Brooks S. Markov chain Monte Carlo method and its application. The Statistician. 1998; 47:69–100.

48. Beaumont MA, Rannala B. The Bayesian revolution in genetics. Nature Reviews Genetics. 2004; 5:251–261. https://doi.org/10.1038/nrg1318 PMID: 15131649

49. Wilkinson DJ. Bayesian methods in bioinformatics and computational systems biology. Briefings in Bioinformatics. 2007; 8:109–116. https://doi.org/10.1093/bib/bbm007 PMID: 17430978

50. Liu JS, Chen R. Sequential Monte Carlo methods for dynamic systems. Journal of the American Statistical Association. 1998; 93:1032–1044. https://doi.org/10.1080/01621459.1998.10473765

51. Chopin N. A sequential particle filter method for static models. Biometrika. 2002; 89:539–552. https://doi.org/10.1093/biomet/89.3.539

52. Doucet A, De Freitas N, Gordon N. An introduction to sequential Monte Carlo methods. In: Sequential Monte Carlo methods in practice. Springer; 2001. p. 3–14.

53. Murphy KP. Machine learning: A probabilistic perspective. MIT Press; 2012.

54. Tikhonov AN, Goncharsky AV, Stepanov VV, Yagola AG. Numerical methods for the solution of ill-posed problems. vol. 328. Springer Science & Business Media; 2013.

55. Kuntsevich A, Kappel F. SolvOpt: The solver for local nonlinear optimization problems. Institute for Mathematics, Karl-Franzens University of Graz. 1997.

56. Tierney L, Kadane JB. Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association. 1986; 81:82–86. https://doi.org/10.1080/01621459.1986.10478240

57. Long Q, Scavino M, Tempone R, Wang S. Fast estimation of expected information gains for Bayesian experimental designs based on Laplace approximations. Computer Methods in Applied Mechanics and Engineering. 2013; 259:24–39. https://doi.org/10.1016/j.cma.2013.02.017

58. Wang K, Bui-Thanh T, Ghattas O. A randomized maximum a posteriori method for posterior sampling of high dimensional nonlinear Bayesian inverse problems. SIAM Journal on Scientific Computing. 2018; 40:A142–A171. https://doi.org/10.1137/16M1060625

59. Konakli K, Sudret B. Reliability analysis of high-dimensional models using low-rank tensor approximations. Probabilistic Engineering Mechanics. 2016; 46:18–36. https://doi.org/10.1016/j.probengmech.2016.08.002

60. Deman G, Konakli K, Sudret B, Kerrou J, Perrochet P, Benabderrahmane H. Using sparse polynomial chaos expansions for the global sensitivity analysis of groundwater lifetime expectancy in a multi-layered hydrogeological model. Reliability Engineering & System Safety. 2016; 147:156–169. https://doi.org/10.1016/j.ress.2015.11.005

61. Lataniotis C, Marelli S, Sudret B. Extending classical surrogate modelling to ultrahigh dimensional problems through supervised dimensionality reduction: A data-driven approach. arXiv preprint arXiv:181206309. 2018.

62. Hampton J, Doostan A. Coherence motivated sampling and convergence analysis of least squares polynomial chaos regression. Computer Methods in Applied Mechanics and Engineering. 2015; 290:73–97. https://doi.org/10.1016/j.cma.2015.02.006

63. Rodriguez-Fernandez M, Egea JA, Banga JR. Novel metaheuristic for parameter estimation in nonlinear dynamic biological systems. BMC Bioinformatics. 2006; 7(1):483. https://doi.org/10.1186/1471-2105-7-483 PMID: 17081289

64. Raue A, Kreutz C, Maiwald T, Bachmann J, Schilling M, Klingmüller U, et al. Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. Bioinformatics. 2009; 25:1923–1929. https://doi.org/10.1093/bioinformatics/btp358 PMID: 19505944

65. Chis O, Banga JR, Balsa-Canto E. Structural identifiability of systems biology models: A critical comparison of methods. PloS One. 2011; 6:e27755. https://doi.org/10.1371/journal.pone.0027755 PMID: 22132135

66. Septier F, Peters GW. An overview of recent advances in Monte Carlo methods for Bayesian filtering in high-dimensional spaces. In: Theoretical Aspects of Spatial-Temporal Modeling. Springer; 2015. p. 31–61.

67. Snyder C, Bengtsson T, Bickel P, Anderson J. Obstacles to high-dimensional particle filtering. Monthly Weather Review. 2008; 136:4629–4640. https://doi.org/10.1175/2008MWR2529.1

68. Gelfand AE, Sahu SK. Identifiability, improper priors, and Gibbs sampling for generalized linear models. Journal of the American Statistical Association. 1999; 94:247–253. https://doi.org/10.1080/01621459.1999.10473840

69. Raue A, Kreutz C, Theis FJ, Timmer J. Joining forces of Bayesian and frequentist methodology: A study for inference in the presence of non-identifiability. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences. 2013; 371:20110544. https://doi.org/10.1098/rsta.2011.0544

70. Atkinson AC, Donev AN. Optimum experimental designs. New York: Oxford University Press; 2007.

71. Mesbah A, Streif S. A probabilistic approach to robust optimal experiment design with chance constraints. IFAC-PapersOnLine. 2015; 48(8):100–105. https://doi.org/10.1016/j.ifacol.2015.08.164

72. Chaloner K, Verdinelli I. Bayesian experimental design: A review. Statistical Science. 1995; p. 273–304. https://doi.org/10.1214/ss/1177009939

73. Ryan EG, Drovandi CC, McGree JM, Pettitt AN. A review of modern computational algorithms for Bayesian optimal design. International Statistical Review. 2016; 84:128–154. https://doi.org/10.1111/insr.12107

74. Paulson JA, Martin-Casas M, Mesbah A. Optimal Bayesian experiment design for nonlinear dynamic systems with chance constraints. Journal of Process Control. 2019; https://doi.org/10.1016/j.jprocont.2019.01.010.

75. Huan X, Marzouk Y. Simulation-based optimal Bayesian experimental design for nonlinear systems. Journal of Computational Physics. 2013; 232:288–317. https://doi.org/10.1016/j.jcp.2012.08.013