Title

Data-driven flow-map models for data-efficient discovery of dynamics and fast uncertainty quantification of biological and biochemical systems.

Journal

Biotechnology and bioengineering, 120(3)

ISSN

0006-3592

Authors

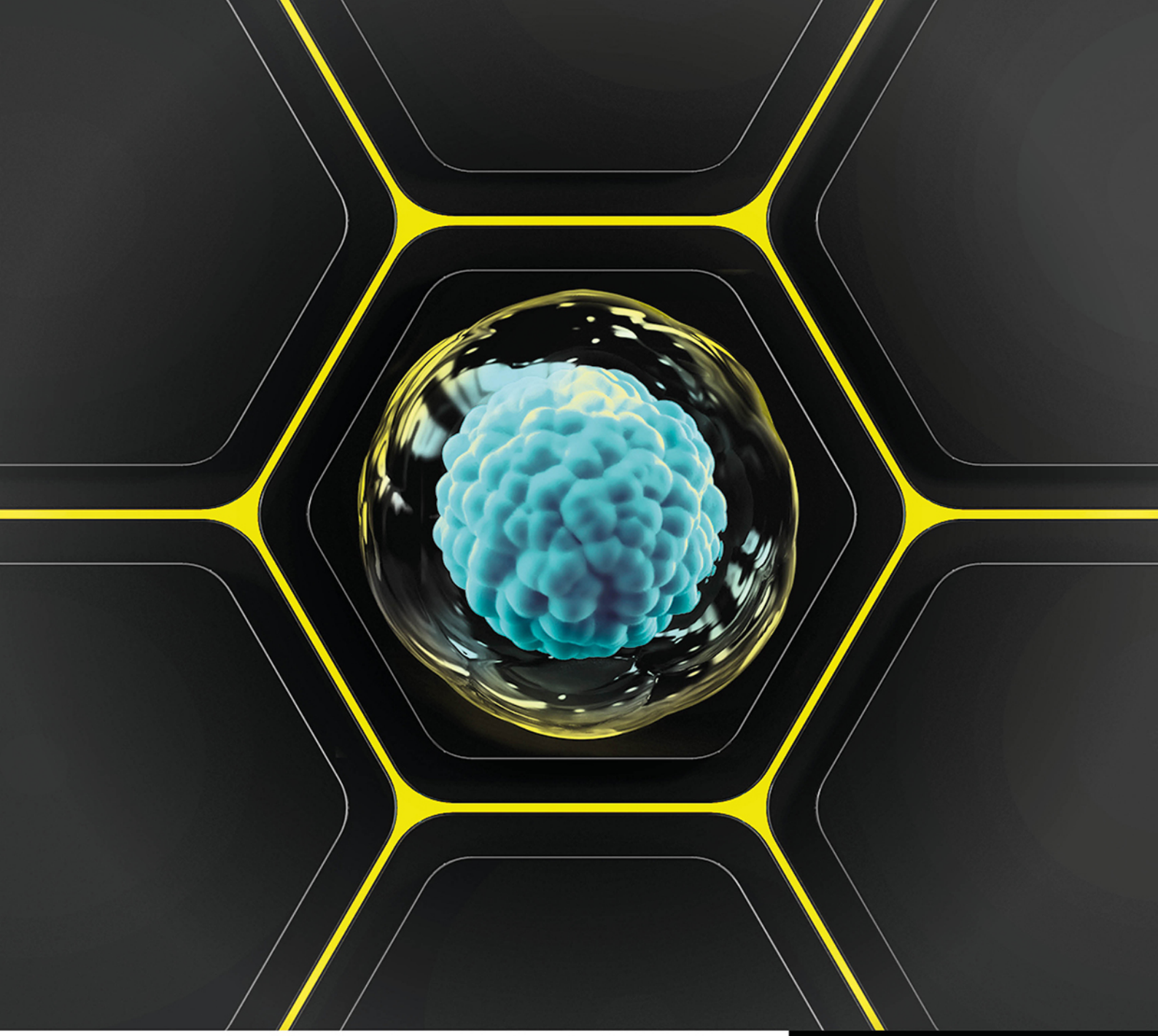
Makrygiorgos, Georgios
Berliner, Aaron J
Shi, Fengzhe
et al.

Publication Date

2023-03-01

DOI

10.1002/bit.28295



Transforming Liquid Biopsy Analysis

Simplifying Progress

Isolate pure single tumor cells from liquid biopsies with CellCelector Flex

Learn how CellCelector Flex can reduce steps, streamline workflows, and accelerate processing - while capturing cells without loss or damage, enabling high viability and up to 100% efficiency.

It's a unique and complete solution for detecting and isolating CTCs after enrichment and staining steps. In just 10 minutes, you can pick 20 individual CTCs for DNA, RNA or proteome analysis.

Learn more:
www.sartorius.com/cellcelector

SARTORIUS

Received: 26 February 2022 | Revised: 27 July 2022 | Accepted: 9 October 2022

DOI: 10.1002/bit.28295

ARTICLE

BIOTECHNOLOGY AND BIOENGINEERING WILEY

Data-driven flow-map models for data-efficient discovery of dynamics and fast uncertainty quantification of biological and biochemical systems

Georgios Makrygiorgos[1,2] | Aaron J. Berliner[1,3] | Fengzhe Shi[1,2] | Douglas S. Clark[1,2] | Adam P. Arkin[1,3] | Ali Mesbah[1,2]

[1]Center for the Utilization of Biological Engineering in Space (CUBES), Berkeley, California, USA

[2]Department of Chemical and Biomolecular Engineering, University of California, Berkeley, California, USA

[3]Department of Bioengineering, University of California, Berkeley, California, USA

Correspondence

Ali Mesbah, Department of Chemical and Biomolecular Engineering, University of California, Berkeley, California, USA.
Email: mesbah@berkeley.edu

Abstract

Computational models are increasingly used to investigate and predict the complex dynamics of biological and biochemical systems. Nevertheless, governing equations of a biochemical system may not be (fully) known, which would necessitate learning the system dynamics directly from, often limited and noisy, observed data. On the other hand, when expensive models are available, systematic and efficient quantification of the effects of model uncertainties on quantities of interest can be an arduous task. This paper leverages the notion of flow-map (de)compositions to present a framework that can address both of these challenges via learning data-driven models useful for capturing the dynamical behavior of biochemical systems. Data-driven flow-map models seek to directly learn the integration operators of the governing differential equations in a black-box manner, irrespective of structure of the underlying equations. As such, they can serve as a flexible approach for deriving fast-to-evaluate surrogates for expensive computational models of system dynamics, or, alternatively, for reconstructing the long-term system dynamics via experimental observations. We present a data-efficient approach to data-driven flow-map modeling based on polynomial chaos Kriging. The approach is demonstrated for discovery of the dynamics of various benchmark systems and a coculture bioreactor subject to external forcing, as well as for uncertainty quantification of a microbial electrosynthesis reactor. Such data-driven models and analyses of dynamical systems can be paramount in the design and optimization of bioprocesses and integrated biomanufacturing systems.

KEYWORDS

discovery of nonlinear dynamics, flow-map decomposition, polynomial chaos Kriging, probabilistic surrogate modeling, uncertainty quantification

1 | INTRODUCTION

Computational models have become indispensable tools for understanding the complex behavior of biological and biochemical systems toward design and optimization of bioprocesses and integrated biomanufacturing systems (Banga et al., 2005). Recently, there has been a growing interest in data-driven methods for modeling the uncertain and nonlinear dynamics of biochemical systems, as these models constitute the cornerstone of various model-based analyses and decision-making tasks such as experiment design, hypothesis

testing and parameter inference (Franceschini & Macchietto, 2008; Golightly & Wilkinson, 2011; Iooss & Lemaître, 2015). Data-driven modeling is especially useful when it is formidable to derive first-principles descriptions for systems whose complex behavior can span over multiple length- and time-scales. Data-driven models have shown promise for inferring the dynamics of cellular systems and metabolic networks (e.g., Daniels & Nemenman, 2015; Schmidt et al., 2011). Hybrid models (aka gray-box models) that combine physics-based models with data-driven descriptions of unknown or hard-to-model phenomena have also proven useful for describing the complex behavior of biochemical systems (De Azevedo et al., 1997; Schubert et al., 1994; VonStosch et al., 2014; Zhang et al., 2019). In this work, we focus on *data-driven discovery* of dynamical systems, whereby the goal is to learn directly the governing equations from system observations. A class of data-driven discovery methods for unknown systems relies on basic assumptions about the structure of the underlying equations (Bongard & Lipson, 2007). To this end, a popular technique is based on sparse identification from dictionaries of possible governing terms (Brunton et al., 2016; Champion et al., 2019), which has been shown to be particularly useful when limited system observations are available. On the other hand, non-parametric modeling approaches relax the necessity of using a library of candidate terms (Heinonen et al., 2018). Another class of methods for data-driven reconstruction of dynamics is based on dynamic mode decomposition (Kutz et al., 2016; Schmid, 2010), which approximates the eigenvalues and eigenvectors of the Koopman operator (Williams et al., 2015) that describes the dynamics of nonlinear systems.

Although inception of the field of nonlinear system identification dates back to few decades ago (Schoukens & Ljung, 2019), the advent of machine learning, in particular deep learning, for characterizing complex input–output relationships has reinvigorated the interest in this area. Most notably, physics-informed neural networks (Raissi et al., 2019) and dynamics reconstruction via neural networks under noisy data (Rudy et al., 2019) have shown promise for data-driven modeling of nonlinear dynamical systems. Recently, Qin et al. (2020, 2019) proposed a deep learning-based approach for data-driven approximation of the integration operator of differential equations from observations of state variables. The usefulness of this approach for discovery of dynamics of biological systems has been demonstrated on several benchmark problems in Su et al. (2021), mainly since it removes the necessity of assumptions about the dynamic model structure.

Data-driven discovery methods can also be used for model-based uncertainty quantification (UQ) applications that rely on expensive-to-evaluate computational models. Predictions of the behavior of biochemical systems are generally subject to various sources of uncertainty due to unknown model structure, parameters, and/or initial and boundary conditions. Systematic and accurate quantification of the effects of these uncertainties on predictions of quantities of interest is crucial when using models for decision-support tasks. This has spurred development of a plethora of set-based (Streif et al., 2016) and probabilistic (Najm, 2009; Smith, 2013) methods for forward and inverse UQ problems (e.g., Komorowski et al., 2009; Mesbah & Streif, 2015; Paulson et al., 2019a; Rumschinski et al., 2010; Vanlier et al., 2013). However, the most commonly used UQ methods rely on Monte Carlo sampling (Caflisch, 1998), which can be intractable for expensive computational models of biochemical systems, especially when models consist of a large number of differential equations and/or have a large number of uncertain inputs.

Surrogate modeling is increasingly used to facilitate complex UQ analyses that would otherwise be computationally prohibitive. The key notion in surrogate modeling is to construct a data-driven mapping between inputs to a system and the quantities of interest in a nonintrusive manner, in which the "data-generating process," for example, a high-fidelity model, is treated as a black-box to generate as few training samples as possible (Sudret et al., 2017). Such a data-driven representation can be used as a computationally efficient surrogate for expensive computational models to predict the output quantities as a function of inputs. A variety of surrogate modeling techniques such a generalized and sparse polynomial chaos (Blatman & Sudret, 2011; Xiu & Karniadakis, 2003), Kriging (Cressie, 1990), and deep learning (Tripathy & Bilionis, 2018) have been successfully applied to various biological and biochemical systems (e.g., delRio-Chanona et al., 2019; Paulson et al., 2019a; Pereira et al., 2021; Schillings et al., 2015; Streif et al., 2014). Nonetheless, a critical challenge in the majority of these techniques arises from capturing the time-evolution of the states in an efficient manner. The most common approach, known as *time-frozen* surrogate modeling (Makrygiorgos et al., 2020; Pettit & Beran, 2006), for predicting the time-evolution of states relies on constructing separate surrogate models for all time points at which the states must be predicted. As such, the "time-frozen" approach can be an inflexible and inefficient way of surrogate modeling for dynamical systems, especially in dynamic UQ and decision-making problems that hinge on making predictions over an adaptive sequence of time instants.

In this paper, we leverage the notion of flow-map (de)composition, as also investigated in Qin et al. (2020, 2019), for data-efficient discovery of system dynamics from experimental observations or high-fidelity simulation data. Conceptually, a flow-map is an analytical operator that maps the current state and input of a system to a future state based on exact integration of model equations over some specified time step. Numerical integration schemes for ordinary differential equations in fact seek to numerically approximate flow-maps to compute the time-evolution of state variables as a function of input variables. Here, we propose to approximate flow-maps in a data-driven manner via nonintrusive surrogate modeling, such that the resulting *data-driven flow-map* is a surrogate for integration operators of the differential equations governing a dynamical system. Hence, data-driven flow-map models are able to discover system dynamics irrespective of the unknown structure of model equations. In addition, data-driven flow-map models can address the above-described challenge of "time-frozen" approaches to surrogate modeling via circumventing the need for construction of separate surrogate models at different time instants. This can be especially

useful for fast UQ and optimization-based analyses of dynamical systems that hinge on repeated runs of expensive computational models over a sequence of time instants.

We demonstrate the usefulness of data-driven flow-maps for discovery of system dynamics from data, as well as for fast UQ applications based on expensive computational models. In this work, sparse polynomial chaos Kriging (PCK; Schöbi & Sudret, 2014) is used for data-driven approximation of flow-maps owing to its data efficiency, ability to approximate complex mappings and ability to quantify the uncertainty of model predictions. The versatility of data-driven flow-maps is first demonstrated via the discovery of the transient behavior of benchmark problems and a coculture bioreactor using noisy data. Subsequently, we show how data-driven flow-maps can speed up forward and inverse UQ analyses of a dynamic microbial electrosynthesis reactor, achieving up to a 100-fold gain in computational speed.

2 | METHODS

In this section, we present the idea of flow-map (de)composition for dynamical nonlinear systems. We first introduce the notion of flow-map functions, which we seek to approximate in a data-driven manner based on time-evolution of system states. This is followed by a discussion on the data generation strategy and the PCK method used in this work to approximate flow-map functions for the variables of interest.

2.1 | Flow-map compositions

Consider a dynamical, time-invariant, nonlinear system described by

$$\frac{d\boldsymbol{s}}{dt} = \boldsymbol{f}(\boldsymbol{s}, \boldsymbol{x}), \quad \boldsymbol{s}(t = 0) = \boldsymbol{s}_0, \tag{1}$$

where $\boldsymbol{s} \in \mathbb{R}^{n_s}$ is the vector of state variables with initial conditions $\boldsymbol{s}_0$, $\boldsymbol{x} \in \mathbb{R}^{n_x}$ is the vector of input variables, and $\boldsymbol{f}(\boldsymbol{s}, \boldsymbol{x}) : \mathbb{R}^{n_s} \times \mathbb{R}^{n_x} \to \mathbb{R}^{n_s}$ is the vector of (possibly unknown) system equations; $\mathbb{R}$ denotes the set of real numbers. Equation (1) describes the time-evolution of the states $\boldsymbol{s}$ of a nonlinear system as a function of the inputs $\boldsymbol{x}$. Notice that in this work the inputs $\boldsymbol{x}$ can represent either model parameters, or manipulated input variables to a biochemical system, as will be discussed later.

A flow-map function is a mapping that predicts the transition of a dynamical system from the current to future state (Qin et al., 2019). We define a flow-map function Φ_δ as

$$\boldsymbol{s}(t + \delta; \boldsymbol{x}) = \Phi_\delta(\boldsymbol{s}_t, \boldsymbol{x}), \tag{2}$$

$$\boldsymbol{s}(t + \delta; \boldsymbol{x}) = \boldsymbol{s}(t; \boldsymbol{x}) + \int_t^{t+\delta} \boldsymbol{f}(\boldsymbol{s}(t'; \boldsymbol{x}), \boldsymbol{x}) dt', \tag{3}$$

where δ is a time-lag (i.e., integration time step). Equation (3) describes the one-step transition between the states of a system in some interval $(t, t + \delta)$. The integral term that appears in Equation (3) can be considered as a flow-map residual since it represents the discrepancy between the current and future states. Although Equation (1) provides a continuous-time description of a dynamical system, the notion of transitioning among states, as implied by Equation (3), hinges on discretizing the time domain over which the system evolves. Accordingly, the idea of flow-map compositions can be applied to compose a sequence of one-step transitions to define state trajectories over time (Qin et al., 2019). Once a sequence of flow-maps $\{\Phi_{\delta_1}, \Phi_{\delta_2}, \ldots, \Phi_{\delta_K}\}$ is established, the flow-maps can be used to predict the states $\boldsymbol{s}$ at any discrete time instant using the K-fold composition

$$\Phi_\Delta = \Phi_{\delta_K} \circ \cdots \circ \Phi_{\delta_1}, \tag{4}$$

where $\circ$ denotes the function composition operator and Δ is the sum of the time-lags over the K discrete time steps (i.e., $\Delta = \sum_{j=1}^{K} \delta_j$). Equation (4) indicates that, starting from some initial states, the K-fold flow-map function Φ_Δ governs the state transitions over the time-lags $\delta_1, \ldots, \delta_K$ wherein at each discrete time step the states are a function of the previous states as given by Φ_{δ_j}. Note that, in general, the time lags δ_j in Equation (4) need not be the same.

In practice, the set of differential equations in Equation (1) describing the system dynamics may not be known, or, when known, their numerical solution may be expensive. In this paper, we aim to learn an approximate surrogate for the flow-map function Φ_δ in Equation (2) from high-fidelity simulation or experimental data. Data-driven flow-map models can be established from simulation data to provide an efficient surrogate for expensive computational models of the form in Equation (1) that, for example, rely on numerical integration of a large number of highly nonlinear and stiff differential equations, as is commonly the case for complex biochemical systems. Notice that in this case data-driven flow-map models essentially approximate a numerical integrator of the differential equations in Equation (1). Alternatively, in the absence of any knowledge about the governing equations (i.e., functions $\boldsymbol{f}$ in Equation 1), flow-map models can be directly learned from experimental observations to discover the unknown system dynamics.

2.2 | Data generation

The data generation and model training strategy adopted in this work is summarized in Figure 1. Consider that we have access to one or more, in total N_T, trajectories of state variables $\boldsymbol{s}_k$ over a discrete-time horizon $k = 0, 1, \ldots, T - 1$, where k is the discrete-time index and T is the length of the time horizon of the training trajectories. Note that $k = 0$ corresponds to t_0 (i.e., the initial time), $k = 1$ to $t_1 = t_0 + \delta_1$, $k = 2$ to $t_2 = t_0 + \delta_1 + \delta_2$, and so forth. The state trajectories can be generated either from simulations or experiments. For some time interval indexed by k and corresponding time t_k where the states are known, we observe a transition in states $\boldsymbol{s}_k \to \boldsymbol{s}_{k+1}$. Hence, given the current states (at time t_k), we obtain the future states (at time

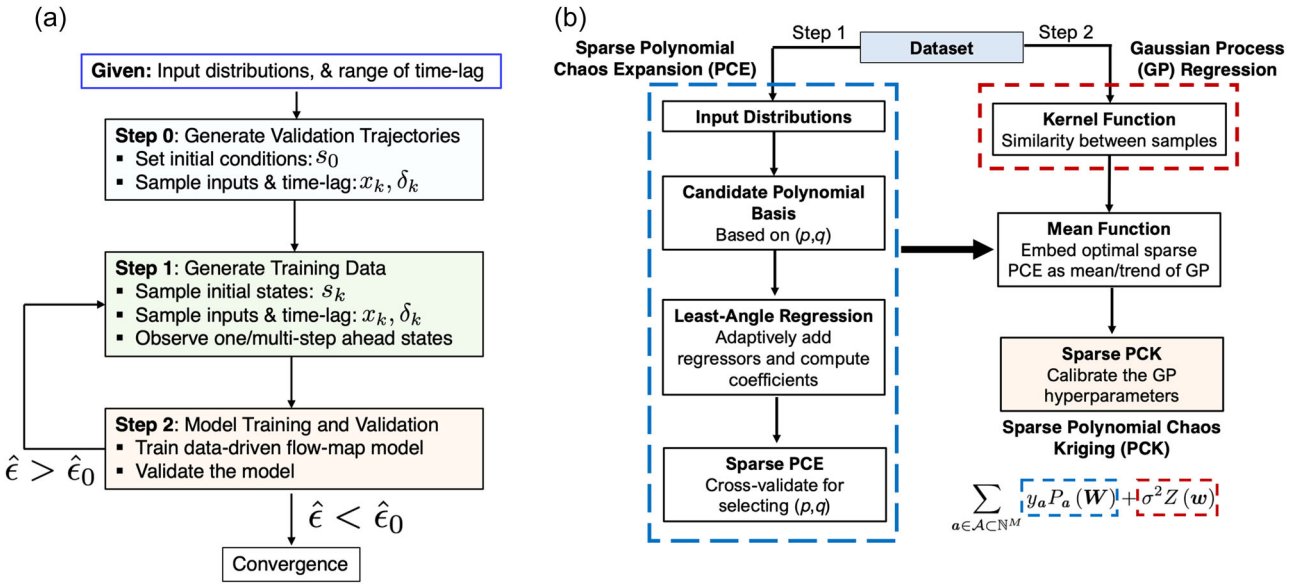


FIGURE 1 (a) Algorithm for data generation and training of data-driven flow-map models. Validation trajectories are first generated. Then, one/multi-step ahead simulations or experiments are performed to observe successor states given the initial states, inputs, and time-lag. Subsequently, the data-driven flow-map model is trained. Finally, the prediction accuracy of the trained model is assessed against the long-time validation trajectories. If the prediction accuracy $\hat{\epsilon}$ is larger than some prespecified threshold $\hat{\epsilon}_0$, the model training and validation process must be repeated. The training procedure for PCK is depicted in (b). Several parameters must be selected during the model training, including the polynomial order, hyperbolic truncation parameter, covariance function, and the regression method used for estimating the expansion coefficients.

$t_k + \delta_k$), while the interval length represents the time lag δ_k. Moreover, within each interval, inputs $\boldsymbol{x}_k$, such as manipulated inputs to the system, may be varying and thus should be accounted for in the data collection procedure. In summary, each time interval in a trajectory contains the information for a one-step transition and, therefore, yields a single sample for the data set.

Accordingly, at each time instant k, the current states s_k, the system input variables of interest x_k, and the lag time δ_k constitute the input variables in the data set, while the corresponding label is the set of states that the system arrives at, that is, s_{k+1}. Thus, the data set (or experimental design) takes the following form: the input data are $\{(\boldsymbol{s}_0, \boldsymbol{x}_0, \delta_0), (\boldsymbol{s}_1, \boldsymbol{x}_1, \delta_1), \ldots, (\boldsymbol{s}_{T-1}, \boldsymbol{x}_{T-1}, \delta_{T-1})\}$ and the corresponding outputs are $\{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_T\}$. Note that there is usually some degree of freedom in choosing the lag time δ in simulations, whereas the choice of δ is often limited by how fast measurements can be acquired in experiments. For trajectory generation, it is crucial to vary the initial conditions $\boldsymbol{s}_0$ and inputs $\boldsymbol{x}_k$ within some allowable range, as well as the time lag δ whenever applicable. The training data must cover a wide range of state, input and time lag values, as relevant to the application of the trained models. We note that an effective strategy for generating simulation data is via one-step transitions, that is, trajectories of length equal to 1. This implies that, instead of generating an entire trajectory given some initial conditions $\boldsymbol{s}_0$, we can randomly sample the state-space, along with the input parameters and time lag, to compute the corresponding future states. The sampling step (Step 1) is summarized in Figure 1a. We remark that, although random sampling is used here to generate the training data, probabilistic models such as PCK used in this work provide confidence estimates on their predictions that can be used towards active learning-based sampling (e.g., see Tsymbalov et al., 2018).

2.3 | Data-driven flow-maps using polynomial chaos kriging

In this work, we use sparse PCK (Makrygiorgos et al., 2020; Schöbi & Sudret, 2014) to discover a data-driven flow-map model $\widetilde{\Phi}(\boldsymbol{w}_k)$ for the dynamical system in Equation (1), i.e., Step 2 in Figure 1a. Note that since the time-lag is part of the inputs, we drop the subscript δ from the flow-map function for notational convenience. The PCK training is summarized in Figure 1b. Let us denote the vector of current states, input variables, and lag time by $\boldsymbol{w}_k = [\boldsymbol{s}_k^\top \boldsymbol{x}_k^\top \delta_k]^\top \in \mathbb{R}^M$, where $M = n_s + n_x + 1$. Thus, we denote the data-driven approximation of the flow-map in Equation (2) by $\widetilde{\Phi}(\boldsymbol{s}_k, \boldsymbol{x}_k, \delta_k) : \mathbb{R}^{n_s} \times \mathbb{R}^{n_x} \times \mathbb{R} \to \mathbb{R}^{n_s}$. The main benefits of using PCK for constructing data-driven flow-map models include: (i) being more data efficient, especially as compared to data intensive feedforward neural networks (Su et al., 2021), when used for discovery of biological system dynamics from system observations; (ii) offering significant improvements in the computational efficiency of data generation for surrogate modeling for dynamical systems as compared to time-frozen polynomial chaos approaches (Mai & Sudret, 2017; Makrygiorgos et al., 2020); and (iii) characterizing the uncertainty of model predictions. To this end, PCK combines the global approximation capability of polynomial

chaos expansions, extensively used for surrogate modeling of (bio) chemical systems (e.g., Deman et al., 2016; Oladyshkin & Nowak, 2012; Paulson et al., 2019b), with the local interpolation scheme of Kriging (i.e., Gaussian processes [GP], Williams & Rasmussen, 2006). The polynomial structure of PCK makes its training data efficient, whereas Kriging offers the ability to quantify the uncertainties of model predictions.

In the context of PCK, $\boldsymbol{w}_k$ is a realization of the multivariate random variable W with a (known) joint probability distribution f_W, that is, $W \sim f_W$. The PCK approximation of the flow-map is defined as

$$\mathcal{Y} = \widetilde{\Phi}(\boldsymbol{w}) = \sum_{\boldsymbol{a} \in \mathcal{A} \subset \mathbb{N}^M} \gamma_{\boldsymbol{a}} P_{\boldsymbol{a}}(\boldsymbol{W}) + \sigma^2 Z(\boldsymbol{w}), \tag{5}$$

where $\mathcal{Y} \in \mathbb{R}^{n_s}$ denotes the predicted variables of interest at step $k+1$, which are typically a subset of the states $\boldsymbol{s}$. The first term in Equation (5) describes the trend (or mean) of the GP using a polynomial chaos expansion (PCE), while the second term $Z(\boldsymbol{w})$ describes the variance of the predicted variable. $P_{\boldsymbol{a}}(\boldsymbol{W})$ represents the multivariate polynomial basis functions that are orthogonal with respect to the probability distribution f_W over the support $\mathcal{D}_W$ of the distribution, that is, the range over which random numbers are defined and can be drawn from. Based on the distribution of the inputs $\boldsymbol{w}$, there is an optimal (or almost optimal) choice for the family of polynomials that are used to guarantee the expansion convergence, either for well-known (Cameron & Martin, 1947; Xiu & Karniadakis, 2002) or for arbitrary (Paulson et al., 2017) distributions; more details about this procedure can be found in the Supporting Information. $\gamma_{\boldsymbol{a}}$ are the coefficients of the basis functions, with the multi-index $\boldsymbol{a}$ being a vector in the set $\mathcal{A}$, which is a subset of natural numbers $\mathbb{N}^M$. The latter notation represents a vector consisting of M elements that are all natural numbers. Therefore, the multi-index here is essentially an extended index that represents the order of each monomial that participates in each polynomial term in the PCE.

Note that originally the multi-index represents any combination of polynomials of any arbitrary order if $\mathcal{A} = \mathbb{N}^M$. Nevertheless, to keep the expansion of the trend term in Equation (5) finite and tractable, it must be truncated up to a finite order p. Furthermore, sparsity in the expansion can be introduced by employing the hyperbolic truncation scheme (Blatman & Sudret, 2011), also known as the q-norm scheme, which removes terms from the expansion that involve the interaction of high order monomials based on some parameter q. Therefore, the allowed values of the multi-index are determined by the tuple (p, q); more details are given in the Supporting Information. In addition, for the GP-related term, $Z(\boldsymbol{w})$ is a standard normal random distribution with variance σ^2.

As described, the multivariate random variable $\boldsymbol{W}$ consists of the states $\boldsymbol{s}$, input variables $\boldsymbol{x}$, and time lag δ. When $\boldsymbol{x}$ corresponds to uncertainties of a computational model (e.g., uncertainties in model parameters and/or boundary conditions), their probability distribution is typically available a priori from parameter inference. As such, their respective polynomial basis functions can be chosen according to the Wiener–Askey scheme (e.g., Hermite basis for Gaussian distributions, Legendre for uniform distributions). On the other hand, when $\boldsymbol{x}$ corresponds to manipulated variables of a system, as is the case in the discovery of system dynamics, the input variables can typically be modeled as uniform distributions within a known range. The time lag δ can also be modeled as a uniform distribution within some range of interest for the application at hand. However, the distribution of states s_k is dependent on the realized state trajectories when the training data are generated and, thus, cannot be established a priori. Here, we assume states follow a multivariate Gaussian distribution with a mean and covariance computed from the training samples.

The coefficients $\gamma_{\boldsymbol{a}}$ of the polynomial chaos expansion can be determined in a nonintrusive manner via solving a least-squares problem (Bishop, 2006). Here, we induce further sparsity by modifying the coefficient estimation problem to a L_1-regularized regression problem (Hastie et al., 2015). The regularized coefficient estimation problem can be efficiently solved using the least-angle-regression (LAR) algorithm (Efron et al., 2004), which estimates the coefficients of the most relevant terms of the expansion, setting the rest of the coefficients to zero. Moreover, $Z(\boldsymbol{w})$ in Equation (5) is defined in terms of a kernel function $R(|\boldsymbol{w} - \boldsymbol{w}'|, \theta)$, that is, a function that provides some measure of similarity between different realizations of the random variable $\boldsymbol{W}$. Here, we use the Matérn kernel function (Williams & Rasmussen, 2006). Overall, the parameters of the PCK that must be determined using the training data include the coefficients $\gamma_{\boldsymbol{a}}$ of the trend, the variance term σ^2, and the hyperparameters θ of the kernel function. The variance and the hyperparameters can be efficiently estimated via maximum-likelihood estimation (Schöbi & Sudret, 2014).

In this work, the following procedure is used for deriving the PCK flow-maps using the data generation scheme of Section 2.2. We use the sequential PCK approach proposed in Schöbi and Sudret (2014), where a PCE is first trained based on the available data and is then embedded as the trend of PCK. This procedure is shown in Figure 1b. For training the PCE, we allow the PCE's maximum order to vary from 1 to 5; higher order polynomials are avoided to retain a smaller expansion (i.e., less degrees of freedom) and mitigate overfitting. The truncation factor q is varied from 0.7 to 0.85 since the resulting maximum order of the polynomials will ensure that we do not have highly nonlinear interaction terms while allowing for elimination of few of interaction terms. The optimal value of q is chosen based on cross-validation. The hyperparameters of PCK are selected using a data-driven optimization algorithm, namely the covariance matrix adaptation-evolution strategy (Hansen & Ostermeier, 2001). Finally, it should be noted that using PCK as the surrogate model places some limitation on the number of input variables $\boldsymbol{w}$ that can be handled. Typically, GP-based models are utilized for lower dimensional spaces due to the "curse-of-dimensionality" (Tripathy et al., 2016).[1] On the other hand, sparse PCEs can effectively deal with high input dimensions thanks not only to the truncation schemes that are employed, but also the sparse regression schemes, for example, LAR, that include only the most informative terms in the expansion, thus minimizing the number of unknown coefficients. To quantify the quality of the PCK predictions during the training, we use the leave-one-out cross-validation (LOOCV) error that is estimated from the

training data (during LAR in Figure 1). When one-step ahead test samples are available, validation errors can readily be evaluated and are used for cross-validation.

Above, we described the flow-map modeling procedure by utilizing one-step transition data, which implies that the PCK model is able to predict one step ahead states. Nevertheless, we are interested in long-term integration of the dynamical system. To this end, Figure 2 shows how a data-driven flow-map model can be used sequentially to predict the time-evolution of the states of a dynamical system. As stated, at each time instant k, the PCK flow-map model essentially "integrates" the states forward in time by δ_k until the final time is reached. As such, we can also assess the ability of the data-driven flow-map models in approximating the integration operator and, hence, their predictive accuracy over a multistep integration horizon. Given $i = 1, \ldots, N_V$ validation state trajectories, each of which of length T_i, we define the normalized, time-averaged prediction error of the state variables, ϵ_i, as

$$\epsilon_i = \sum_{k=0}^{T_i} \frac{1}{T_i} \frac{\left\| \mathcal{Y}_{k,i} - \mathcal{Y}_{k,i}^{\text{true}} \right\|_2}{\left\| \mathcal{Y}_{k,i}^{\text{true}} \right\|_2}, \tag{6a}$$

$$\hat{\epsilon} = \frac{1}{N_V} \sum_{i=1}^{N_V} \epsilon_i, \tag{6b}$$

where $\| \cdot \|_2$ is the 2-norm of a vector; $\mathcal{Y}_{k,i}^{\text{true}}$ and $\mathcal{Y}_{k,i}$ are, respectively, the vector of state variables in the validation data set and those predicted by the data-driven flow-map models at time instant k for each validation run i. In the remainder, we refer to ϵ_i as the mean trajectory error (MTE), whereas $\hat{\epsilon}$ is the average MTE over all validation trajectories.

3 | DATA-DRIVEN DISCOVERY OF DYNAMICAL SYSTEMS

In this section, we apply the PCK-based flow-map modeling approach to learn the dynamics of several benchmark systems using limited data. The first case study, based on the Morris-Lecar system, compares the performance of the PCK model with neural network modeling results of Su et al. (2021). The second case study, based on the Lorenz system, focuses on reconstructing the dynamics of a chaotic system in which variations in parameters significantly change the solution landscape. Lastly, we show how the flow-map modeling approach can be used for discovering the dynamics of a coculture bioreactor under noisy observations and how the variance term of PCK provides a measure of uncertainty of model predictions.

3.1 | Morris–Lecar system

The first benchmark problem is the Morris–Lecar system proposed by Morris and Lecar (1981), which describes neuronal excitability. This system was used in Su et al. (2021) to examine neural network-based flow-map models for the discovery of nonlinear dynamics. In particular, a residual neural network was used to represent the data-driven flow-map model, in which only the flow-map residual is learned by skipping the input connection to the neural network and adding it to the output of the latter. Here, we aim to recreate the results of the aforementioned work, demonstrating the data efficiency of the proposed PCK approach for data-driven reconstruction of dynamics. The dynamics of the Morris–Lecar system are described by

$$C_M \frac{dV}{dt} = -g_L(V - V_L) - g_{Ca}(V - V_{Ca})M_\infty - g_k(V - V_K)N + I_{\text{app}} \tag{7a}$$

$$\frac{dN}{dt} = \lambda_N(N_\infty - N), \tag{7b}$$

where V (mV) is the voltage difference between the sides of the membrane and N represents the probability for the potassium channel being open. The parameters M_∞, N_∞, and λ_N depend on the voltage, as defined in the Supporting Information. We focus on the so-called Type I model with parameters taken from Su et al. (2021) and given in the Supporting Information. Here, it is assumed that the model parameters are fixed since we aim to reconstruct the system dynamics as a function of injected current $x_k = I_{\text{app}}$ that can vary within the range [0, 300] A. Specifically, we aim to predict the long-term system dynamics, starting from given initial conditions, under a fixed I_{app}. To compare our results with those in Su et al. (2021), δ_k was chosen to be 0.2 ms; we did not consider the time-lag as part of the PCK model. In other words, the input data consisted of samples in the form $(V_k, N_k, x_k) \rightarrow (V_{k+1}, N_{k+1})$, where x_k is constant for every $k \in [0, T]$ for a given trajectory. This system exhibits a saddle node bifurcation, which leads to an oscillatory behavior depending on the value of input I_{app}. Thus, the data-driven flow-map model must capture the oscillatory behavior of the Morris–Lecar system for different values of I_{app}.

To train the PCK-based flow-map model, we generated one-step ahead samples of the states V_k and N_k by randomly drawing the initial

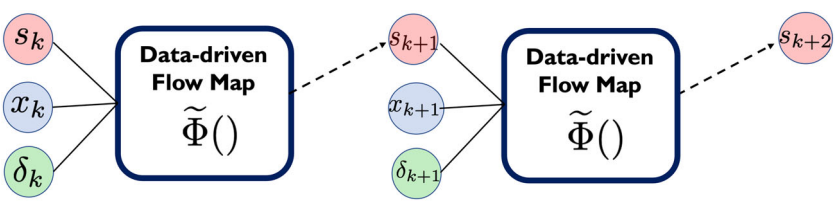


FIGURE 2 Data-driven flow-map models for predicting the state variables of a dynamical system over time. The flow-map model $\widetilde{\Phi}$ takes the current states $\boldsymbol{s}_k$, inputs $\boldsymbol{x}_k$, and lag time δ_k at a discrete-time instant k as inputs to predict the states $\boldsymbol{s}_{k+1}$ at the subsequent time instant $k + 1$. By sequentially repeating this procedure, the time-evolution of the states in relation to the inputs can be established.

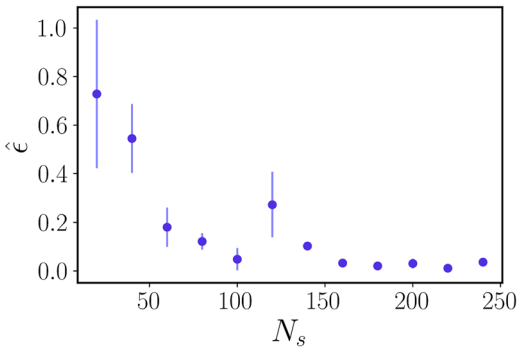


FIGURE 3 The average mean trajectory error, ê, of the PCK-based flow-map model for the Morris–Lecar system in relation to the number of training samples, N_s. The error is estimated based on three validation trajectories generated for the input I_{app} values {0, 60, 150}. The vertical bars represent the standard deviation of the error estimated based on five repeats of the training.

states from [−75, 75] × [0, 1]. Here, we first examine the convergence error of the flow-map model to characterize how many samples of states would be necessary for data-driven reconstruction of the system dynamics. We quantify the convergence error in terms of the average MTE in Equation (6a) based on three validation trajectories generated for I_{app} = {0, 60, 150}. Figure 3 shows the average MTE estimated over 1000 time steps in relation to the number of training samples, where the vertical bar around each error represents one standard deviation based on five repetitions of the analysis. It is evident that the error converges after about 160 samples, suggesting that a limited number of training samples is needed.

Figure 4 shows the reconstructed dynamics by the PCK-based flow-map model trained using 240 samples in comparison with the true dynamics. As can be seen, there is no visible discrepancy between the true time-evolution of the system and the reconstructed dynamics. The system exhibits a bifurcation behavior, as evident from

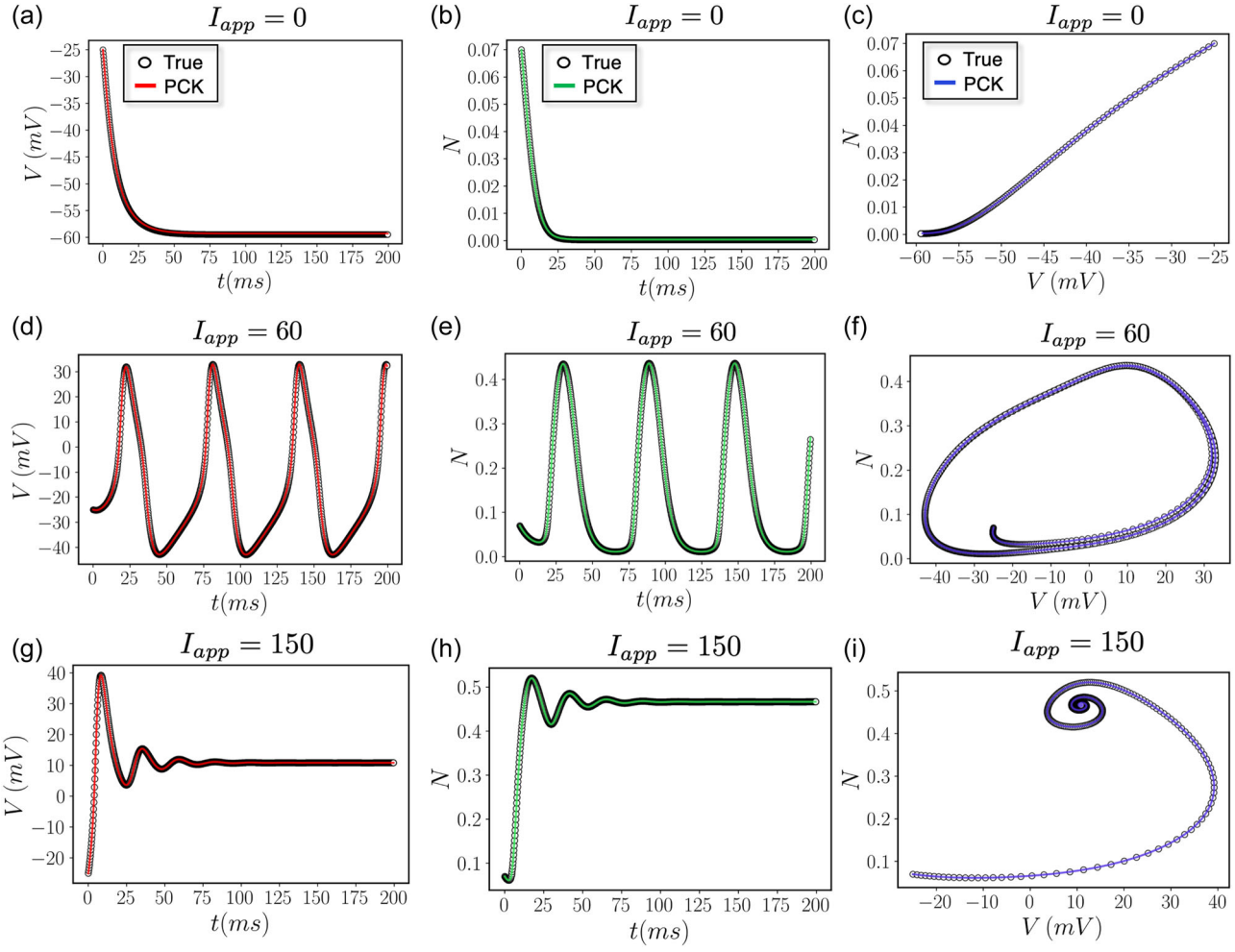


FIGURE 4 Reconstructed dynamics of the Morris–Lecar system by the PCK-based flow-map model in comparison with the true system dynamics for the input I_{app} values {0, 60, 150}. The PCK-based flow-map model is trained using 240 samples. The left column shows the time-evolution of voltage difference, V; the middle column shows the time-evolution of the channel opening probability, N; and the right column shows the corresponding phase plots.

the phase plots shown in Figure 4c,f,i. Yet, the PCK-based flow-map model is able to capture this complex behavior and accurately predict the system dynamics over a long-time horizon. We note that a 500-fold saving in the number of training samples is observed as compared to Su et al. (2021) in which a residual neural network representation was used for the flow-map model. This is while the PCK model also yields slightly more accurate predictions.

3.2 | Lorenz system

We now consider a chaotic dynamical system based on the well-known Lorenz benchmark problem presented in Sparrow (2012). The Lorenz system has been widely used in the data-driven modeling literature (e.g., Dubois et al., 2020; Raissi et al., 2018). The Lorenz system is described by the following set of nonlinear ordinary differential equations

$$\frac{da}{dt} = \sigma(b - a), \tag{8a}$$

$$\frac{db}{dt} = a(\rho - c) - b, \tag{8b}$$

$$\frac{dc}{dt} = ab - \beta c, \tag{8c}$$

where $\boldsymbol{s} = [a, b, c]^\top$ are the system states and $\boldsymbol{x} = [\sigma, \rho, \beta]^\top$ are the uncertain model parameters. Chaotic behaviors can be encountered in various chemical and biological systems, including in the growth of biological populations with nonoverlapping generations (May, 1974) and the peroxidase-oxidase oscillator (Olsen & Lunding, 2021). Here, we consider a constant time-lag $\delta = 0.01$ that captures the intrinsic time-scale of the system (Brunton & Kutz, 2019).

The Lorenz system exhibits a chaotic behavior based on the initial conditions $\boldsymbol{s}_0$, while its long-term behavior is highly affected by the uncertain parameters $\boldsymbol{x}$. The nominal initial conditions and parameters of the system are, respectively, $\boldsymbol{s}_0 = [1.9427, -1.4045, 0.9684]^\top$ and $\boldsymbol{x}_0 = [10, 28, 8/3]^\top$, for which the system oscillates around two attractors. Here, the training data consisted of 500 random samples of the states $\boldsymbol{s}$ within the range $[-10, 10] \times [-10, 10] \times [-10, 10]$ and the parameters $\boldsymbol{x}$ within the range $[8, 12] \times [10, 30] \times [1, 5.5]$. We used two validation trajectories to compare the true system dynamics with those reconstructed by the PCK-based flow-map model: one trajectory based on the nominal initial conditions and parameters and the other based on $\boldsymbol{x} = [10, 15, 8/3]^\top$ and $\boldsymbol{s}_0 = [1.6655, -0.1178, 0.1748]^\top$.

Figure 5 shows phase plots of the reconstructed oscillatory dynamics of the Lorenz system in comparison with the true system dynamics over a simulation horizon of 5000 time steps. We observe that the qualitative behavior of the Lorenz system is different when the parameter ρ is varied, while the PCK-based flow-map model is able to reconstruct the dynamics in both cases. The MTE is 0.522 for the nominal validation trajectory and 0.0013 for the second validation trajectory. Although the error for the nominal validation trajectory seems relatively high, the main characteristics of the true dynamics are adequately captured, as evident from Figure 5a–c. That is, the limit cycles, the amplitude of oscillation and period are adequately captured. These predictions are consistent with those reported in Raissi et al. (2018). However, we note that reconstruction

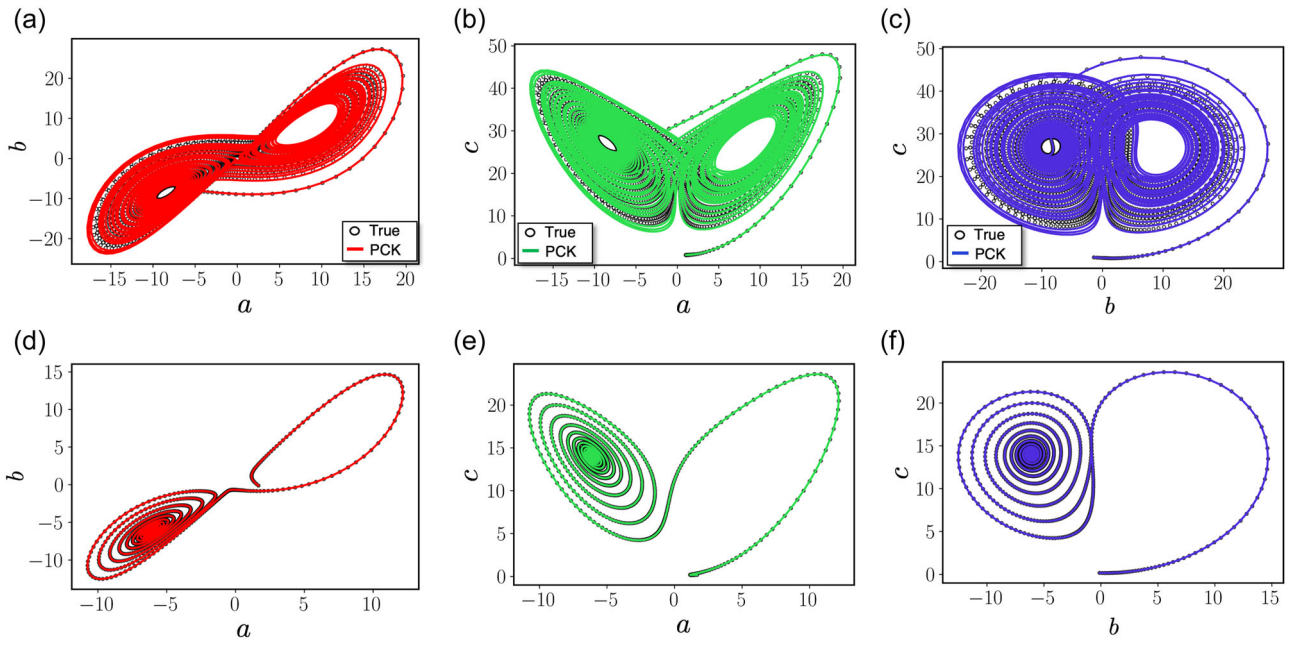


FIGURE 5 Phase plots of the reconstructed dynamics of the Lorenz system by the PCK-based flow-map model in comparison with the true system dynamics for different values of model parameters. Subplots (a)–(c) correspond to the model parameters $\sigma = 10$, $\beta = 8/3$, and $\rho = 28$. Subplots (d)–(f) correspond to the model parameters $\sigma = 10$, $\beta = 8/3$, and $\rho = 15$.

of the Lorenz dynamics using neural networks typically requires on the order of a few thousands of training samples (Brunton & Kutz, 2019; Rudy et al., 2019), whereas the PCK model here was trained using 500 samples.

3.3 | Transient coculture system

In this case study, we demonstrate the ability of PCK-based flow-map models to learn the transient behavior of a coculture system with variable inputs. In particular, we focus on the startup dynamics of a continuous bioreactor driven by the competition of several auxotrophs (Pande et al., 2014). To emulate data collection from a real system, we use a nonlinear dynamic model of the bioreactor (Treloar et al., 2020) to generate observations of the system states, which are then corrupted with independent and identically distributed state-dependent measurement noise $e_i \sim \mathcal{N}(0, 2.5 \times 10^{-2} s_k^i)$, with i being an index for the measured states and k the time index. The five state variables $\boldsymbol{s}_k$ of the bioreactor include: the population of the two species N_1 (Cells/L) and N_2 (Cells/L), the auxotrophic nutrients concentrations C_1 (g/L) and C_2 (g/L), and the common shared carbon source concentration C_0 (g/L). The bioreactor has three process inputs $\boldsymbol{x}_k$ that can be varied in time. The process inputs are the dilution rate D (h^{-1}) that varies within the range [0.75, 1.5] (h^{-1}), as well as the feed substrate concentration of auxotrophs $C_{1,in}$ (g/l) and $C_{2,in}$, both varying in the range [1.5, 2] (g/l). To generate data for training the PCK-based flow map models, short simulation "experiments" with a fixed length of $T = 30$ steps with $\delta_k \in [0.15, 0.25] h^{-1}$ were performed. At each time step k during the multistep experiments, inputs $\boldsymbol{x}_k$ were varied over the time interval δ_k and noisy observations of the states were collected.

Figure 6 shows the state trajectories for a validation set versus those predicted by the PCK-based flow map models. The validation trajectories were generated by some random initial conditions at $k = 0$ and applying an input x_0 over the interval δ_0. The model predicts the mean of the states at $k = 1$, as well as their variance. The integration proceeds by taking a next step based on the mean value of the states at $k = 1$, predicting the states at $k = 2$, and so forth. Using only the mean value to compute trajectories is the simplest way when Gaussian process regression models are utilized, however, there are more sophisticated ways for the trajectory generation (Hewing et al., 2020), which are beyond the scope of the paper. Note that the predicted uncertainty information can also be incorporated into multistep ahead predictions, as discussed in Girard et al., 2003 and Polymenakos et al., 2017. Here, it suffices to use a deterministic function, for example, the mean value of the data-driven flow-map model, to integrate in time since this way we avoid the major issue of using noisy inputs into the PCK model. In Figure 6, the validation trajectories have a length of $N_k = 40$ steps, extending slightly beyond the training range. Moreover, we can characterize the confidence in

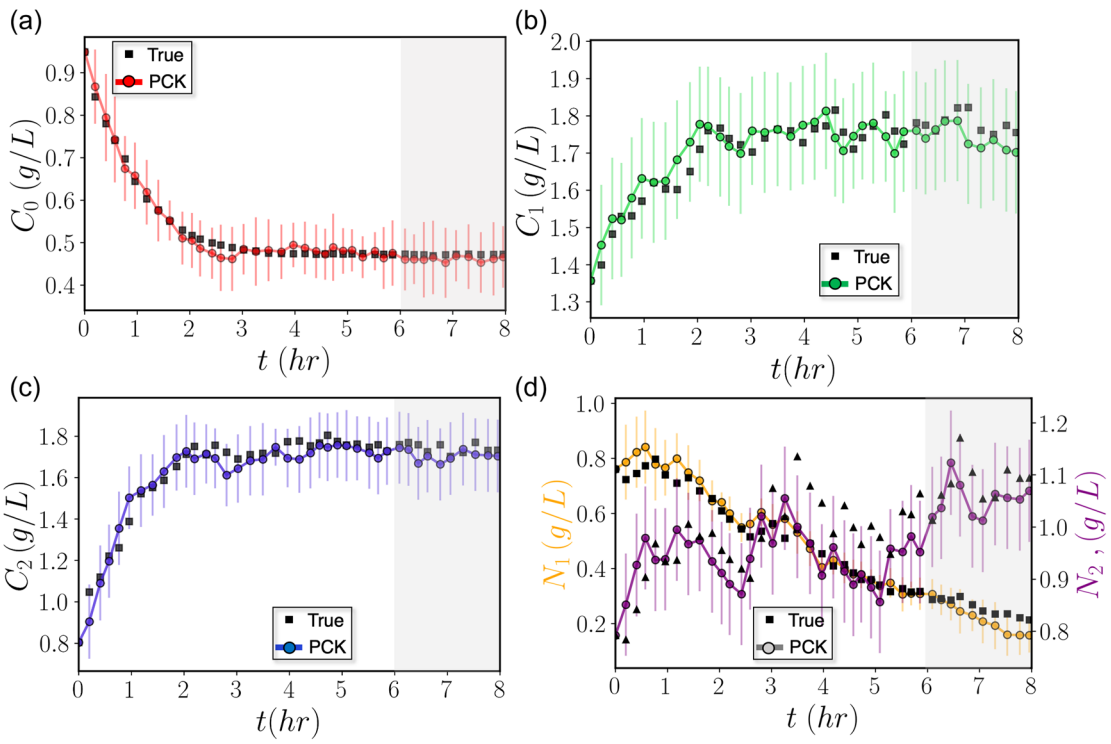


FIGURE 6 Predictions of the state variables of the transient coculture system via the PCK-based flow-map models in comparison with the observed state trajectories. The colored lines/points correspond to the predicted trajectories by the mean of the PCK models, starting from some initial states at $t = 0$ hr. Black symbols represent the observed trajectories at specific snapshots during a validation run. Vertical error bars represent the uncertainty in the predictions of the PCK models, estimated as plus/minus three standard deviations from the mean value. The shaded areas correspond to a time interval that was not accounted for when training the PCK models.

the predictions of the dynamics. To this end, at each step k, we plot the plus/minus three standard deviation error bars around the predicted mean of each state variable. As can be seen, the true state trajectories are within the confidence intervals of the PCK predictions.

4 | UNCERTAINTY QUANTIFICATION OF EXPENSIVE COMPUTATIONAL MODELS

In this section, we demonstrate the utility of data-driven flow-maps for the UQ of a microbial electrosynthesis (MES) bioreactor using a high-fidelity computational model that is subject to uncertainty in model parameters and initial conditions. In particular, we show how flow-maps can be used as surrogate models for efficient sample-based approximation of distribution of state variables, global sensitivity analysis, and Bayesian parameter inference when the original model is prohibitively expensive for a sample-based analysis.

We consider the batch MES bioreactor shown in Figure 7 for CO_2 fixation (Abel & Clark, 2021), with potential applications in space biomanufacturing (Berliner et al., 2021). The bioreactor consists of a well-mixed liquid bulk phase that contains dissolved CO_2, that is, substrate. A microbial community forming a biofilm grows on the cathode of the bioreactor. The dissolved substrate diffuses into the biofilm through a linear boundary layer and is then consumed by bacteria toward the growth of the biofilm. This leads to spatial distribution of the substrate concentration within the biofilm. Voltage is applied to the cathode while the biofilm acts as a conductive matrix through which electron transport takes place. Both the substrate CO_2 in the biofilm and the local overpotential due to the current flux contribute to the biofilm growth kinetics described by the dual Monod–Nerst model (Torres et al., 2008).

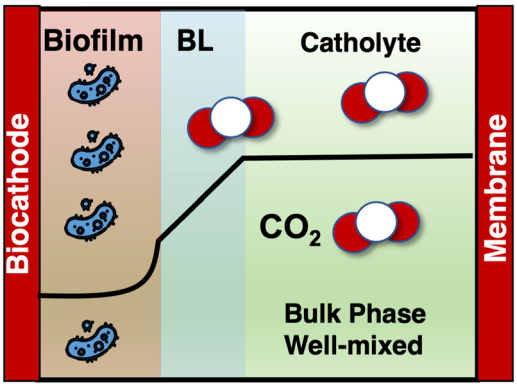


FIGURE 7 Schematic of the microbial electrosynthesis bioreactor. The bioreactor consists of three regions: the bulk phase, the biofilm, and a boundary layer (BL) in between. The black line represents a typical concentration profile of some species as predicted by the computational model used in this work. The concentration is assumed to be constant in the bulk phase, changing linearly across the boundary layer, and exhibiting a more complicated shape in the biofilm.

A computational model of the dynamics of the MES bioreactor is adopted from Kazemi et al. (2015) and Marcus et al. (2007), with some modifications. Within the biofilm, the cell growth leads to the production of acetate as a metabolic product. A primary modeling approach in the aforementioned papers assumes the total biomass has a constant concentration and exists in two forms, active and inactive, each of which occupies some volume fraction. We assume that biomass exists only in active form, thus the equations describing the volume-fraction change within the film effectively become a single equation for the rate of change of film thickness, L_f, which is a differential state in our system. Moreover, the film growth is affected by a constant detachment rate. It is also assumed that the reaction occurs only within the biofilm, so the only source of acetate in the bulk phase comes from exchange with the biofilm through the boundary layer. We further assume the transport-reaction phenomena in the biofilm are much faster than the transport that occurs across the boundary layer and in the bulk phase. Accordingly, the conservation laws inside the biofilm are considered to be in pseudo steady-state (Kazemi et al., 2015). Hence, the computational model consists of a set of nonlinear second-order ordinary differential equations that describe the spatial distribution of substrate, acetate and overpotential within the biofilm, coupled with a set of first-order ordinary differential equations that describe the concentration of CO_2 in the bulk phase S_b, the acetate concentration in the bulk phase P_b, and the biofilm thickness L_f. As such, the three state variables of the system are described by

$$\frac{dL_f}{dt} = (Y\hat{q} - r_d)L_f, \tag{9a}$$

$$\frac{dS_b}{dt} = \frac{A_f}{V_r} j_S, \tag{9b}$$

$$\frac{dP_b}{dt} = \frac{A_f}{V_r} j_P, \tag{9c}$$

where $Y\left(\frac{\text{mgX}}{\text{mmolS}}\right)$ is the biomass yield coefficient, $\hat{q}\left(\frac{\text{mmolS}}{\text{mgXdays}}\right)$ represents an average substrate consumption specific rate within the biofilm, $r_d\left(\frac{1}{\text{days}}\right)$ is a detachment rate, A_f (cm^2) is the cross-sectional area of the biofilm, and V_r (cm^3) is the bioreactor volume. The mass balances for the substrate and product are a function of the flux of each species across the linear boundary layer as described by

$$j_m = \frac{D_b}{L_b}(m_f(z = L_f) - m_b), \quad m = S, P, \tag{10}$$

where m denotes the species (i.e., substrate and product), $D_b\left(\frac{\text{cm}^2}{\text{days}}\right)$ is the diffusivity coefficient in the boundary layer and L_b (cm) is the thickness of the boundary layer. The subscript f denotes the species concentration in the film at position $z = L_f$. The equations that describe the diffusion phenomena within the film are given in the Supporting Information. To determine the concentrations at L_f, a boundary value problem (diffusion within the film) must be solved at each time step, as

the concentrations in the biofilm are a function of the bulk concentrations. The computational model is fairly expensive for UQ analyses that rely on Monte Carlo sampling; each model run takes on average 4 min. The model is also subject to time-invariant uncertainty in its parameters and initial conditions. Specifically, the model uncertainty comprises of the conductivity of the biofilm k_{bio}, the maximum growth rate μ_{max} of the Nerst–Monod model, the yield Y, the Monod affinity constant K_s, as well as the acetate production-related parameters α and β. These six uncertain parameters are assumed to follow a uniform probability distribution. Their nominal values are $[k_{bio}, \mu_{max}, Y, K_s, \alpha, \beta]^T = [1 \times 10^{-3}, 4.5, 0.25, 3.0, 0.1, 2 \times 10^{-5}]^T$, while they vary uniformly ±10% about their nominal values.

In this case study, we construct data-driven flow-map models of the PCK form in Equation (5) for the output variables $\mathcal{Y} = [L_f\ S_b\ P_b]^T$, such that the six sources of uncertainty constitute the vector of input variables $\boldsymbol{x}$ in Equation (5). The three flow-map models, one for each state variable, were trained using simulation data generated via the computational model for lag times in the range of $\delta = [0.05, 0.1]$ days to allow us to adequately capture the bioreactor dynamics. Notice that clearly the lag time δ must always be larger than the integration time step of the computational model.

The training data set consists of full state trajectories, as well as one-step ahead samples of the states. We initially generate $N_T = 30$ trajectories, with fixed uncertain parameters in time, over a process time span from 0 to 3.5 days that corresponds to approximately $T = 50$ samples per trajectory. Then, using the states $\boldsymbol{s}_k$ corresponding to each sample $\boldsymbol{w}_k$, we randomize the uncertain parameters and perform one-step ahead simulations. In this way, approximately 1400 training samples were generated, while 800 samples are used for training the PCK models. The rationale behind not randomizing the states is that the validation trajectories (Step 0 of Figure 1a) indicate that there is a high correlation among state values. For instance, as L_f grows in time (under insignificant detachment), S_b decreases due to consumption. Thus, for a given set of uncertain parameters and initial states, a few full state trajectories will help generate more informative training samples. Figure 8 shows the predicted trajectories using the data-driven flow-map PCK model for a given realization of uncertainty and initial conditions, while the true trajectory is juxtaposed. The trajectories correspond to a time-march of 50 steps ahead. We observe a perfect agreement between the predicted and validation trajectories, with the average MTE for the three states being approximately $\hat{\epsilon} = 2.5 \times 10^{-4}$.

An important remark should be made here regarding the benefits of the presented flow-map approach to surrogate modeling of dynamical systems in comparison with the so-called time-frozen approaches discussed in Section 1. First, the flow-map models provide the flexibility to approximate the distribution of states at any time instant of interest without the need for constructing a separate surrogate model for each time instant, as in time-frozen surrogate modeling. For example, if we were to use a time-frozen approach, 50 separate PCK models would need to be constructed to predict the time-evolution of the distribution of the state variables over the 50 time instants considered in this problem. Thus, not only a flow-map modeling approach significantly reduces the number of surrogate models that must be constructed to only one model for each state variable, it also provides flexibility via alleviating the need to build the models at prespecified time points. Furthermore, the flow-map modeling approach enables more efficient data generation. To clarify this point, let us assume that N_p realizations of uncertainties are sufficient for generating a rich training data set that yields surrogate models with low approximation error. In the case of the time-frozen approach, we would require to generate N_p full state trajectories since the states must be observed at all time instants for all uncertainty realizations. This approach to data generation can become prohibitively expensive, in particular when data generation relies on expensive simulations. On the other hand, training the flow-map models in principle requires simulation of a limited number of full state trajectories (in this study, 25 trajectories), whereas N_p training samples can be straightforwardly generated via one-step ahead integration of the computational model. In the following, the use of PCK-based flow-map models is demonstrated for expensive UQ analyses.

4.1 | Forward uncertainty propagation and global sensitivity analysis

Here, we use the data-driven flow-map models for efficient uncertainty propagation via sample-based approximation of the distribution of the three state variables. Figure 9a–c shows the

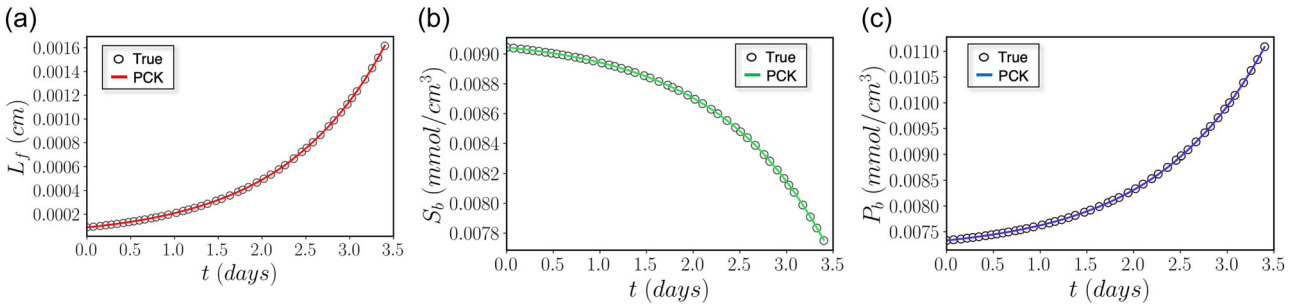


FIGURE 8 Predicted state trajectories of the the microbial electrosynthesis bioreactor: (a) biofilm thickness, L_f, (b) CO_2 concentration in the bulk phase, S_b, and (c) acetate concentration in the bulk phase, P_b. Hollow points represent the validation trajectories, while the solid lines represent the trajectories predicted by the PCK-based flow-map models.

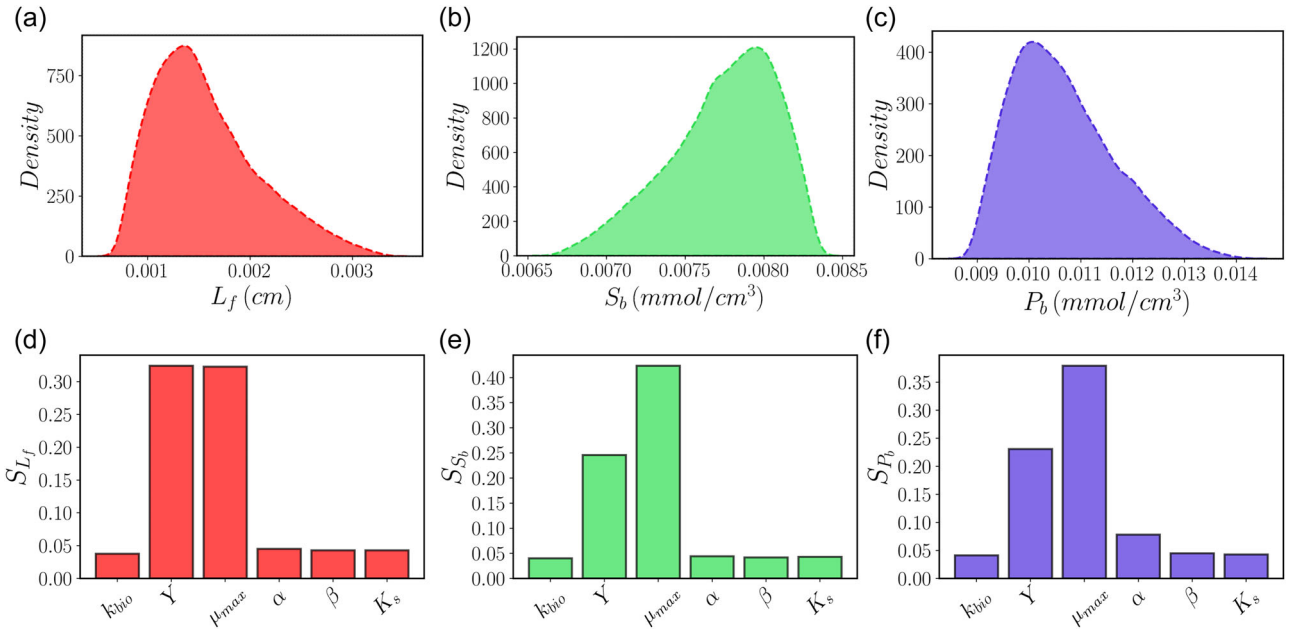


FIGURE 9 Fast uncertainty propagation and global sensitivity analysis of the the microbial electrosynthesis bioreactor using data-driven flow-map models of quantities of interest. Subplots (a)–(c) show the kernel density estimates of the distribution of the biofilm thickness (L_f), concentration of CO_2 in the bulk phase (S_b), and acetate concentration in the bulk phase (P_b) predicted by the PCK models at time t = 3.5 days. The distributions of L_f, S_b, and P_b are approximated via Monte Carlo sampling using 20,000 realizations of uncertain model parameters, where a 100-fold computational speedup in sample-based approximation of the distributions is attained. Subplots (d)–(f) show the Borgonovo indices, denoted by $\mathcal{S}$, that quantify the global sensitivity of L_f, S_b, and P_b at t = 3.5 days with respect to the six uncertain model parameters. The Borgonovo indices are approximated based on 20,000 uncertainty realizations.

distribution of the states at t = 3.5 days. To approximate their distribution, the flow-map models were evaluated using 20,000 realizations of the model uncertainty $\boldsymbol{x}$. Each run of the data-driven flow-map model takes on average less than 3 s,[2] as opposed to the average run time of 4 min of the computational model. This implies that the flow-map models significantly accelerate the uncertainty propagation, enabling an approximately 100-fold increase in the computational speed. This is especially beneficial when the distributions are skewed (or bi-modal), as in Figure 9a-c. In this case, a large number of samples, on the order of 10^4–10^5 samples, would be typically required for accurate sampled-based approximation of the distribution, or statistical moments of the quantities of interest. Although not shown here, we can efficiently approximate the distribution of states at any time instant using trajectories generated by the surrogate models.

Moreover, we use the data-driven flow-map models to perform a global sensitivity analysis to asses the importance of the six uncertain model parameters, $\boldsymbol{x}$, on the state variables $\mathcal{Y}$. This is done via evaluation of the Borgonovo indices (Borgonovo, 2007), denoted by $\mathcal{S}$, which are based on the full distribution of the state variables, as opposed to their statistical moments. The results of global sensitivity analysis for the states at t = 3.5 days are shown in Figure 9d–f, where each bar corresponds to a different uncertain model parameter. The Borgonovo indices are approximated using the same 20,000 samples used in the forward UQ analysis. We observe that the probabilistic uncertainty of yield Y and maximum growth rate $\mu_{\max}$ have the most dominant effects on the variability of the three states, while the product concentration P_b is also significantly affected by the uncertainty in the parameter α, which is the metabolism-related productivity constant.

4.2 | Bayesian inference of unknown model parameters

We now use the data-driven flow-map models to solve a Bayesian inference problem to infer the uncertain model parameters $\boldsymbol{x}$. Bayesian inference relies on Bayes theorem to estimate the posterior probability distribution of the unknown model parameters from available data. Here, noisy observations of L_f, S_b, and P_b at time instants {0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5} days constitute the data set $\mathcal{D}$ used for parameter inference; measurement noise is modeled as a Gaussian distribution with zero mean and state-dependent variance. Once a vector of system measurements $\boldsymbol{d}$ at a time instant is observed, the change in our knowledge about the unknown parameters is described by Bayes' rule (Kennedy & O'hagan, 2001)

$$f_{\mathbf{x}|\mathcal{D}}(\mathbf{x}|\boldsymbol{d}) = \frac{f_{\mathcal{D}|\mathbf{x}}(\boldsymbol{d}|\mathbf{x}) f_{\mathbf{x}}(\mathbf{x})}{f_{\mathcal{D}}(\boldsymbol{d})}, \tag{11}$$

where $f_{\mathbf{x}|\mathcal{D}}$ denotes the posterior distribution of the uncertain parameters after observing the data; $f_{\mathcal{D}|\mathbf{x}}$ is the likelihood function that describes the probability of observing data given the parameter

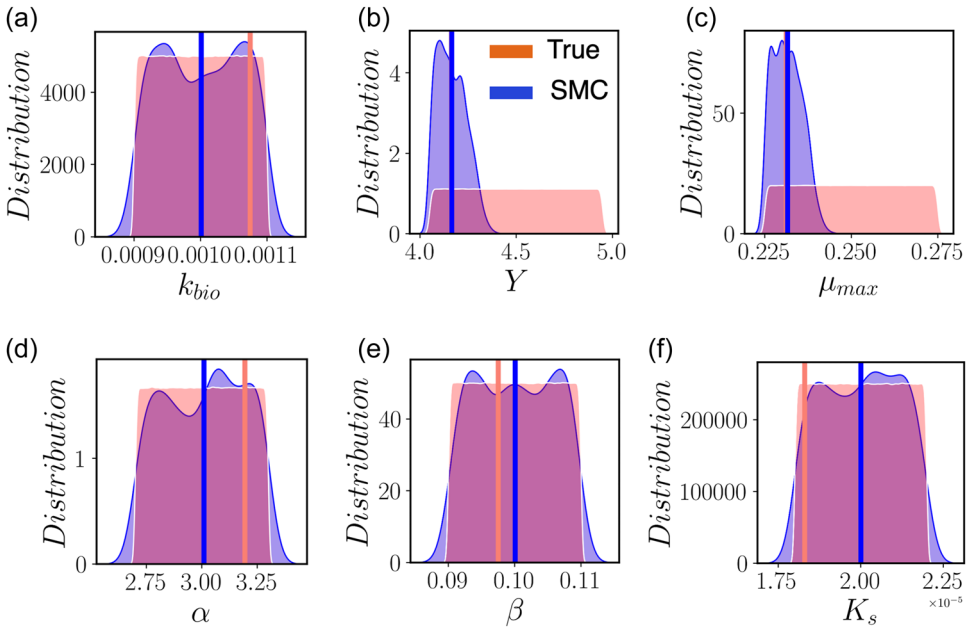


FIGURE 10 Bayesian inference of unknown parameters of the computational model of the microbial electrosynthesis bioreactor. The parameters are estimated via sequential Monte Carlo using 20,000 particles. Red and blue distributions represent the prior and posterior distributions of the unknown model parameters at time 3.5 days, respectively. The red vertical lines correspond to the true parameters, while the blue vertical lines are the estimated posterior mean value of parameters.

estimates; $f_{\boldsymbol{x}}$ is the prior distribution of parameters; and $f_{\mathcal{D}}$ is the so-called evidence or marginal likelihood that ensures the posterior distribution integrates to 1.

As Equation (11) implies, Bayesian inference provides an explicit representation of the uncertainty in the parameter estimates via characterizing the full posterior distribution of unknown parameters $\boldsymbol{x}$. The prior distribution of parameters and the likelihood function must be specified to solve Equation (11). Here, we used the same uniform distributions as those used to construct the PCK surrogate models to represent the prior distributions, although these can be different. The likelihood function is specified by the observation noise model, which is assumed to be zero-mean Gaussian with state-dependent variance in this work. We use a particle filtering method, namely sequential Monte Carlo (SMC) (Liu & Chen, 1998), to approximately solve the Bayesian inference problem by iteratively updating the posterior $f_{\boldsymbol{x}|\mathcal{D}}$ at every time instant that system observations become available; see Makrygiorgos et al. (2020) for further details. Notice that parameter estimation via Bayesian inference methods such as SMC relies on accurate construction of the probability distributions in Equation (11). As described in Section 4.1, the data-driven flow-map models enable efficient sample-based approximation of the distributions using a very large number of samples, which otherwise could be impractical using an expensive computational model.

Figure 10 shows the posterior distribution of the parameters $\boldsymbol{x}$ at t = 3.5 days estimated via SMC using the data set $\mathcal{D}$, as specified above. The posterior distributions are approximated using 20,000 particles. Note that ranges of the posterior distributions are larger than the prior distributions in some cases, which is an artifact of the kernel density estimation (i.e., the selection of the bandwidth parameter; Davis et al., 2011). Figure 10 suggests that only the posterior distributions of parameters Y and $\mu_{\max}$ have changed significantly with respect to their priors. It is also evident that the mean of the posterior distributions (blue vertical lines) for parameters Y and $\mu_{\max}$ provides a fairly accurate estimate for the true, but unknown, parameter values (red vertical lines). In particular, the true value and the posterior mean are indistinguishable, while the posteriors are much more narrow compared to priors as stated before. Nonetheless, the posterior distributions for the other parameters remain similar to their priors with little to no change, suggesting these parameters cannot be estimated using the available data set $\mathcal{D}$. This can be attributed to the lack of information content of system observations $\mathcal{D}$ for inferring the unknown parameters; a deficiency that can be addressed via optimal experiment design (Paulson et al., 2019b; Rodrigues et al., 2020). We again note the flexibility of the flow-map models that would allow us to seamlessly add new observation points, should that become necessary for better parameter inference, without the need to construct new surrogate models for the states observed at new time points.

5 | CONCLUSIONS

This paper presented a flow-map modeling approach based on polynomial chaos Kriging for the discovery of system dynamics from data. Data-driven flow-map models directly approximate the integration operator of differential equations that describe the state transitions of a dynamical system as a function of system state and

input variables. We illustrated the usefulness of the proposed approach for learning mathematical descriptions of nonlinear dynamical systems and deriving dynamic surrogate models for fast uncertainty quantification applications. Our analyses reveal that polynomial chaos Kriging-based flow-maps offer significant benefits in terms of data efficiency, as well as computational efficiency of data generation, for the discovery of nonlinear system dynamics and surrogate modeling.

AUTHORS CONTRIBUTIONS

Georgios Makrygiorgos and Ali Mesbah conceived the concept of this contribution. Georgios Makrygiorgos performed the analysis with help from Aaron J. Berliner and Fengzhe Shi and oversight from Ali Mesbah, Adam P. Arkin, and Douglas S. Clark. Georgios Makrygiorgos, Ali Mesbah, and Aaron J. Berliner wrote the manuscript. All authors edited the manuscript.

ACKNOWLEDGEMENT

This material is based upon work supported by NASA under grant or cooperative agreement award number NNX17AJ31G.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

All software required for reproducing the case studies presented is available through the CUBES github organization at https://github.com/cubes-space/DataDriven-FlowMaps and any additional data is available upon request.

ENDNOTES

[1] Not only the concept of Euclidian distance (which is the main feature of kernels) becomes less meaningful in higher dimensions, but also higher-dimensional input spaces require more data to be efficiently discovered, hence rendering the inference part of GP-based models harder. This challenge can be mitigated via sparse Gaussian process regression (Snelson & Ghahramani, 2007; Titsias, 2009).

[2] Notice that the evaluation time of a PCK model depends on a multitude of factors, such as the degree of the polynomial basis functions, kernel type, and, mainly, the amount of data used to train the model. Additionally, a kernel-based model such as PCK is more expensive to evaluate than a polynomial chaos expansion.

REFERENCES

Abel, A. J., & Clark, D. S. (2021). A comprehensive modeling analysis of formate-mediated microbial electrosynthesis. *ChemSusChem*, *14*, 344–355. https://doi.org/10.1002/cssc.202002079

Banga, J. R., Balsa-Canto, E., Moles, C. G., & Alonso, A. A. (2005). Dynamic optimization of bioprocesses: Efficient and robust numerical strategies. *Journal of Biotechnology*, *117*, 407–419. https://doi.org/10.1016/j.jbiotec.2005.02.013

Berliner, A. J., Hilzinger, J. M., Abel, A. J., McNulty, M. J., Makrygiorgos, G., Averesch, N. J. H., Sen Gupta, S., Benvenuti, A., Caddell, D. F., Cestellos-Blanco, S., Doloman, A., Friedline, S., Ho, D., Gu, W., Hill, A., Kusuma, P., Lipsky, I., Mirkovic, M., … Arkin, A. P. (2021). Towards a biomanufactory on Mars. *Frontiers in Astronomy and Space Sciences*, *8*, 120. https://www.frontiersin.org/article/10.3389/fspas.2021.711550. https://doi.org/10.3389/fspas.2021.711550

Bishop, C. M. (2006). Pattern recognition and machine learning (Vol. 4, pp. 738). Springer.

Blatman, G., & Sudret, B. (2011). Adaptive sparse polynomial chaos expansion based on least angle regression. *Journal of Computational Physics*, *230*, 2345–2367. https://doi.org/10.1016/j.jcp.2010.12.021

Bongard, J., & Lipson, H. (2007). Automated reverse engineering of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, *104*, 9943–9948. https://doi.org/10.1073/pnas.0609476104

Borgonovo, E. (2007). A new uncertainty importance measure. *Reliability Engineering & System Safety*, *92*, 771–784. http://www.sciencedirect.com/science/article/pii/S0951832006000883. https://doi.org/10.1016/j.ress.2006.04.015

Brunton, S. L., & Kutz, J. N. (2019). *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge University Press.

Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. Proceedings of the National Academy of Sciences of the United States of America, *113*, 3932–3937. https://doi.org/10.1073/pnas.1517384113

Caflisch, R. E. (1998). Monte Carlo and quasi-Monte Carlo methods. *Acta Numerica*, *7*, 1–49. https://doi.org/10.1017/S0962492900002804

Cameron, A. R. H., & Martin, W. T. (1947). The orthogonal development of non-linear functionals in series of Fourier-Hermite functionals. *Annals of Mathematics*, *48*, 385–392.

Champion, K., Lusch, B., Kutz, J. N., & Brunton, S. L. (2019). Data-driven discovery of coordinates and governing equations. *Proceedings of the National Academy of Sciences of the United States of America*, *116*, 22445–22451. https://doi.org/10.1073/pnas.1906995116

Cressie, N. (1990). The origins of kriging. *Mathematical Geology*, *22*, 239–252. https://doi.org/10.1007/BF00889887

Daniels, B. C., & Nemenman, I. (2015). Efficient inference of parsimonious phenomenological models of cellular dynamics using s-systems and alternating regression. *PloS ONE*, *10*, e0119821. https://doi.org/10.1371/journal.pone.0119821

Davis, R. A., Lii, K.-S., & Politis, D. N. (2011). Remarks on some nonparametric estimates of a density function. In *Selected works of Murray Rosenblatt* (pp. 95–100). Springer. https://doi.org/10.1007/978-1-4419-8339-8_13

De Azevedo, S. F., Dahm, B., & Oliveira, F. (1997). Hybrid modelling of biochemical processes: A comparison with the conventional approach. *Computers & Chemical Engineering*, *21*, S751–S756. https://doi.org/10.1016/S0098-1354(97)87593-X

delRio-Chanona, E. A., Wagner, J. L., Ali, H., Fiorelli, F., Zhang, D., & Hellgardt, K. (2019). Deep learning-based surrogate modeling and optimization for microalgal biofuel production and photobioreactor design. *AIChE Journal*, *65*, 915–923. https://doi.org/10.1002/aic.16473

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

Dubois, P., Gomez, T., Planckaert, L., & Perret, L. (2020). Data-driven predictions of the Lorenz system. *Physica D: Nonlinear Phenomena*, *408*, 132495. https://doi.org/10.1016/j.physd.2020.132495

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., Ishwaran, H., Knight, K., Loubes, J. M., Massart, P., Madigan, D., Ridgeway, G., Rosset, S., Zhu, J. I., Stine, R. A., Turloptiach, B. A., Weisberg, S., Johnstone, I., &

Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*. *32*, 407–499. https://doi.org/10.1214/009053604000000067

Franceschini, G., & Macchietto, S. (2008). Model-based design of experiments for parameter precision: State of the art. *Chemical Engineering Science*, *63*, 4846–4872. https://doi.org/10.1016/j.ces.2007.11.034

Girard, A., Rasmussen, C., Candela, J. Q., & Murray-Smith, R. (2002). Gaussian process priors with uncertain inputs application to multiple-step ahead time series forecasting. Advances in neural information processing systems, 15.

Golightly, A., & Wilkinson, D. J. (2011). Bayesian parameter inference for stochastic biochemical network models using particle Markov chain Monte Carlo. *Interface Focus*, *1*, 807–820. https://doi.org/10.1098/rsfs.2011.0047

Hansen, N., & Ostermeier, A. (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, *9*, 159–195. https://doi.org/10.1162/106365601750190398

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity: The lasso and generalizations*. Chapman & Hall/CRC.

Heinonen, M., Yildiz, C., Mannerström, H., Intosalmi, J., & Lähdesmäki, H. (2018). Learning unknown ODE models with Gaussian processes. In *International conference on machine learning* (pp. 1959–1968). PMLR. https://proceedings.mlr.press/v80/heinonen18a.html

Hewing, L., Arcari, E., Fröhlich, L. P., & Zeilinger, M. N. (2020). On simulation and trajectory prediction with gaussian process dynamics. In *Learning for Dynamics and Control* (pp. 424–434). PMLR. https://proceedings.mlr.press/v120/hewing20a.html

Iooss, B., & Lemaître, P. (2015). A review on global sensitivity analysis methods. In *Uncertainty management in simulation-optimization of complex systems* (pp. 101–122). Springer. https://doi.org/10.1007/978-1-4899-7547-8_5

Kazemi, M., Biria, D., & Rismani-Yazdi, H. (2015). Modelling bio-electrosynthesis in a reverse microbial fuel cell to produce acetate from CO_2 and H_2O. *Physical Chemistry Chemical Physics 17*, 12561–12574. https://doi.org/10.1039/C5CP00904A

Kennedy, M. C., & O'hagan, A. (2001). Bayesian calibration of computer models. *Journal of Royal Statistical Society B*, *63*, 425–464. https://doi.org/10.1111/1467-9868.00294

Komorowski, M., Finkenstädt, B., Harper, C. V., & Rand, D. A. (2009). Bayesian inference of biochemical kinetic parameters using the linear noise approximation. *BMC Bioinformatics*, *10*, 1–10. https://doi.org/10.1186/1471-2105-10-343

Kutz, J. N., Brunton, S. L., Brunton, B. W., & Proctor, J. L. (2016). *Dynamic mode decomposition: Data-driven modeling of complex systems*. SIAM.

Liu, J. S., & Chen, R. (1998). Sequential Monte Carlo methods for dynamic systems. *Journal of the American Statistical Association*, *93*, 1032–1044. https://doi.org/10.1080/01621459.1998.10473765

Mai, C. V., & Sudret, B. (2017). Surrogate models for oscillatory systems using sparse polynomial chaos expansions and stochastic time warping. *SIAM/ASA Journal on Uncertainty Quantification*, *5*, 540–571.

Makrygiorgos, G., Gupta, S. S., Menezes, A. A., & Mesbah, A. (2020). Fast probabilistic uncertainty quantification and sensitivity analysis of a Mars life support system model. *IFAC-PapersOnLine*, *53*, 7268–7273. https://doi.org/10.1016/j.ifacol.2020.12.563

Makrygiorgos, G., Maggioni, G. M., & Mesbah, A. (2020). Surrogate modeling for fast uncertainty quantification: Application to 2D population balance models. *Computers & Chemical Engineering*, *138*, 106814. https://doi.org/10.1016/j.compchemeng.2020.106814

Marcus, A. K., Torres, C., & Rittmann, B. (2007). Conduction-based modeling of the biofilm anode of a microbial fuel cell. *Biotechnology and Bioengineering*, *98*(6), 1171–1182. https://doi.org/10.1002/bit.21533

May, R. M. (1974). Biological populations with nonoverlapping generations: Stable points, stable cycles, and chaos. *Science*, *186*, 645–647. https://doi.org/10.1126/science.186.4164.645

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

Morris, C., & Lecar, H. (1981). Voltage oscillations in the barnacle giant muscle fiber. *Biophysical Journal*, *35*, 193–213. https://doi.org/10.1016/S0006-3495(81)84782-0

Najm, H. N. (2009). Uncertainty quantification and polynomial chaos techniques in computational fluid dynamics. *Annual Review of Fluid Mechanics*, *41*, 35–52. https://doi.org/10.1146/annurev.fluid.010908.165248

Oladyshkin, S., & Nowak, W. (2012). Data-driven uncertainty quantification using the arbitrary polynomial chaos expansion. *Reliabbility Engineering and System Safety*, *106*, 179–190. https://doi.org/10.1016/j.ress.2012.05.002

Olsen, L. F., & Lunding, A. (2021). Chaos in the peroxidase-oxidase oscillator. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *31*, 013119. https://doi.org/10.1063/5.0022251

Pande, S., Merker, H., Bohl, K., Reichelt, M., Schuster, S., De Figueiredo, L. F., Kaleta, C., & Kost, C. (2014). Fitness and stability of obligate cross-feeding interactions that emerge upon gene loss in bacteria. *The ISME Journal*, *8*, 953–962. https://doi.org/10.1038/ismej.2013.211

Paulson, J. A., Buehler, E. A., & Mesbah, A. (2017). Arbitrary polynomial chaos for uncertainty propagation of correlated random variables in dynamic systems. *IFAC-PapersOnLine*, *50*, 3548–3553. https://doi.org/10.1016/j.ifacol.2017.08.954

Paulson, J. A., Martin-Casas, M., & Mesbah, A. (2019a). Fast uncertainty quantification for dynamic flux balance analysis using non-smooth polynomial chaos expansions. *PLoS Computational Biology*, *15*, e1007308. https://doi.org/10.1371/journal.pcbi.1007308

Paulson, J. A., Martin-Casas, M., & Mesbah, A. (2019b). Optimal Bayesian experiment design for nonlinear dynamic systems with chance constraints. *Journal of Process Control*, *77*, 155–171. https://doi.org/10.1016/j.jprocont.2019.01.010

Pereira, F. H., Schimit, P. H., & Bezerra, F. E. (2021). A deep learning based surrogate model for the parameter identification problem in probabilistic cellular automaton epidemic models. *Computer Methods and Programs in Biomedicine*, *205*, 106078. https://doi.org/10.1016/j.cmpb.2021.106078

Pettit, C., & Beran, P. (2006). Spectral and multiresolution wiener expansions of oscillatory stochastic processes. *Journal of Sound and Vibration*, *294*, 752–779. https://doi.org/10.1016/j.jsv.2005.12.043

Polymenakos, K., Abate, A., & Roberts, S. (2019). Safe policy search using Gaussian process models. In *Proceedings of the 18th international conference on autonomous agents and multiagent systems* (pp. 1565–1573).

Qin, T., Chen, Z., Jakeman, J., & Xiu, D. (2020). Deep learning of parameterized equations with applications to uncertainty quantification. https://doi.org/10.1615/Int.J.UncertaintyQuantification.2020034123

Qin, T., Wu, K., & Xiu, D. (2019). Data driven governing equations approximation using deep neural networks. *Journal of Computational Physics*, *395*, 620–635. https://doi.org/10.1016/j.jcp.2019.06.042

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2018). Multistep neural networks for data-driven discovery of nonlinear dynamical systems. arXiv:1801.01236. https://arxiv.org/abs/1801.01236

Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2018). Multistep neural networks for data-driven discovery of nonlinear dynamical systems.

Rodrigues, D., Makrygiorgos, G., & Mesbah, A. (2020). Tractable global solutions to Bayesian optimal experiment design. In *2020 59th IEEE Conference on Decision and Control* (pp. 1614–1619). IEEE. https://doi.org/10.1109/CDC42340.2020.9304226

Rudy, S. H., Kutz, J. N., & Brunton, S. L. (2019). Deep learning of dynamics and signal-noise decomposition with time-stepping constraints.

Journal of Computational Physics, *396*, 483–506. https://doi.org/10.1016/j.jcp.2019.06.056

Rumschinski, P., Borchers, S., Bosio, S., Weismantel, R., & Findeisen, R. (2010). Set-base dynamical parameter estimation and model invalidation for biochemical reaction networks. *BMC Systems Biology*, *4*, 1–14. https://doi.org/10.1186/1752-0509-4-69

Schillings, C., SunnÅker, M., Stelling, J., & Schwab, C. (2015). Efficient characterization of parametric uncertainty of complex (bio) chemical networks. *PLoS Computational Biology*, *11*, e1004457. https://doi.org/10.1371/journal.pcbi.1004457

Schmid, P. J. (2010). Dynamic mode decomposition of numerical and experimental data. *Journal of Fluid mechanics*, *656*, 5–28. https://doi.org/10.1017/S0022112010001217

Schmidt, M. D., Vallabhajosyula, R. R., Jenkins, J. W., Hood, J. E., Soni, A. S., Wikswo, J. P., & Lipson, H. (2011). Automated refinement and inference of analytical models for metabolic networks. *Physical Biology*, *8*, 055011. https://doi.org/10.1088/1478-3975/8/5/055011

Schöbi, R., & Sudret, B. (2014). PC-Kriging: A new metamodelling method combining polynomial chaos expansions and kriging. Proceedings of the 2nd International Symposium on Uncertainty Quantification and Stochastic Modelling. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.722.6530&rep=rep1&type=pdf

Schoukens, J., & Ljung, L. (2019). Nonlinear system identification: A user-oriented road map. *IEEE Control Systems Magazine*, *39*, 28–99. https://doi.org/10.1109/MCS.2019.2938121

Schubert, J., Simutis, R., Dors, M., Havlik, I., & Lübbert, A. (1994). Bioprocess optimization and control: Application of hybrid modelling. *Journal of Biotechnology*, *35*, 51–68. https://doi.org/10.1016/0168-1656(94)90189-9

Smith, R. C. (2013). *Uncertainty quantification: Theory, implementation, and applications* (vol. 12). SIAM.

Snelson, E., & Ghahramani, Z. (2007). Local and global sparse Gaussian process approximations. In *Artificial Intelligence and Statistics* (pp. 524–531). PMLR.

Sparrow, C. (2012). *The Lorenz equations: Bifurcations, chaos, and strange attractors* (vol. 41). Springer Science & Business Media. https://doi.org/10.1007/978-1-4612-5767-7

Streif, S., Kim, K. K. K., Rumschinski, P., Kishida, M., Shen, D. E., Findeisen, R., & Braatz, R. D. (2016). Robustness analysis, prediction, and estimation for uncertain biochemical networks: An overview. *Journal of Process Control*, *42*, 14–34. https://doi.org/10.1016/j.jprocont.2016.03.004

Streif, S., Petzke, F., Mesbah, A., Findeisen, R., & Braatz, R. D. (2014). Optimal experimental design for probabilistic model discrimination using polynomial chaos. *IFAC Proceedings Volumes*, *47*, 4103–4109. https://doi.org/10.3182/20140824-6-ZA-1003.01562

Su, W.-H., Chou, C.-S., & Xiu, D. (2021). Deep learning of biological models from data: Applications to ODE models. *Bulletin of Mathematical Biology*, *83*, 1–19. https://doi.org/10.1007/s11538-020-00851-7

Sudret, B., Marelli, S., & Wiart, J. (2017). Surrogate models for uncertainty quantification: An overview. In *2017 11th European Conference on Antennas and Propagation* (pp. 793–797). https://doi.org/10.23919/EuCAP.2017.7928679

Titsias, M. (2009). Variational learning of inducing variables in sparse Gaussian processes. In *Artificial intelligence and statistics* (pp. 567–574). PMLR.

Torres, C. I., Marcus, A. K., Parameswaran, P., & Rittmann, B. E. (2008). Kinetic experiments for evaluating the Nernst–Monod model for anode-respiring bacteria (ARB) in a biofilm anode. *Environmental Science & Technology*, *42*, 6593–6597. https://doi.org/10.1021/es800970w

Treloar, N. J., Fedorec, A. J., Ingalls, B., & Barnes, C. P. (2020). Deep reinforcement learning for the control of microbial co-cultures in bioreactors. *PLoS Computational Biology*, *16*, e1007783. https://doi.org/10.1371/journal.pcbi.1007783

Tripathy, R., Bilionis, I., & Gonzalez, M. (2016). Gaussian processes with built-in dimensionality reduction: Applications to high-dimensional uncertainty propagation. *Journal of Computational Physics*, *321*, 191–223.

Tripathy, R. K., & Bilionis, I. (2018). Deep UQ: Learning deep neural network surrogate models for high dimensional uncertainty quantification. *Journal of Computational Physics*, *375*, 565–588. https://doi.org/10.1016/j.jcp.2018.08.036

Tsymbalov, E., Panov, M., & Shapeev, A. (2018). Dropout-based active learning for regression. In *International conference on analysis of images, social networks and texts* (pp. 247–258). Springer.

Vanlier, J., Tiemann, C. A., Hilbers, P. A. J., & Van Riel, N. A. W. (2013). Parameter uncertainty in biochemical models described by ordinary differential equations. *Mathematical Biosciences*, *246*, 305–314. https://doi.org/10.1016/j.mbs.2013.03.006

VonStosch, M., Oliveira, R., Peres, J., & de Azevedo, S. F. (2014). Hybrid semi-parametric modeling in process systems engineering: Past, present and future. *Computers & Chemical Engineering*, *60*, 86–101. https://doi.org/10.1016/j.compchemeng.2013.08.008

Williams, K. I., & Rasmussen, C. (2006). *Gaussian processes for machine learning*. MIT Press. https://doi.org/10.1142/S0129065704001899

Williams, M. O., Kevrekidis, I. G., & Rowley, C. W. (2015). A data-driven approximation of the Koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, *25*, 1307–1346. https://doi.org/10.1007/s00332-015-9258-5

Xiu, D., & Karniadakis, G. E. (2002). The Wiener–Askey polynomial chaos for stochastic differential equations. *SIAM Journal of Science and Computation*, *24*, 619–644. https://doi.org/10.1137/S1064827501387826

Xiu, D., & Karniadakis, G. E. (2003). Modeling uncertainty in flow simulations via generalized polynomial chaos. *Journal of Computational Physics*, *187*, 137–167. https://doi.org/10.1016/S0021-9991(03)00092-5

Zhang, D., DelRio-Chanona, E. A., Petsagkourakis, P., & Wagner, J. (2019). Hybrid physics-based and data-driven modeling for bioprocess online simulation and optimization. *Biotechnology and Bioengineering*, *116*, 2919–2930. https://doi.org/10.1002/bit.27120

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Makrygiorgos, G., Berliner, A. J., Shi, F., Clark, D. S., Arkin, A. P., & Mesbah, A. (2023). Data-driven flow-map models for data-efficient discovery of dynamics and fast uncertainty quantification of biological and biochemical systems. *Biotechnology and Bioengineering*, 120, 803–818. https://doi.org/10.1002/bit.28295