Efficient global sensitivity analysis of biochemical networks using Gaussian process regression

Kurdyaeva, Tamara; Milias-Argeitis, Andreas

Published in:
2018 IEEE Conference on Decision and Control, CDC 2018

DOI:
10.1109/CDC.2018.8618902

2018 IEEE Conference on Decision and Control (CDC)
Miami Beach, FL, USA, Dec. 17-19, 2018

Efficient global sensitivity analysis of biochemical networks using Gaussian process regression

Tamara Kurdyaeva and Andreas Milias-Argeitis

***Abstract*— A key objective of systems biology is to understand how the uncertainty in parameter values affects the responses of biochemical networks. Variance-based sensitivity analysis is a powerful approach to address this question. However, commonly used implementations based on (Quasi-) Monte Carlo require a very large number of model evaluations, and are thus impractical for computationally expensive models. Here, we present an alternative method for variance-based sensitivity analysis that uses Gaussian process regression. Thanks to the appealing mathematical properties of Gaussian processes, we are able to derive exact analytic formulas for the required sensitivity indices. In this way our approach yields more accurate estimates with significantly less computational cost compared to conventional methods, as we demonstrate for a nonlinear model of a bacterial signaling system.**

I. INTRODUCTION

Computational modeling of biochemical networks is usually carried out under large parametric uncertainty due to the inherent difficulties of quantifying parameter values from biological experiments. It is therefore important to know how this uncertainty propagates to outputs of interest and affects the quality of model predictions. Sensitivity analysis methods provide answers to these questions [1]. *Local* sensitivity methods examine the effect of infinitesimal parameter and initial condition perturbations on the model behavior at a relatively small computational cost [1]. However, local sensitivity measures are informative only around a reference point and cannot provide a comprehensive picture of how large parameter and initial condition variations propagate to the outputs. This is the goal of *global* sensitivity analysis [2], which unfortunately also comes at an increased computational cost.

Variance-based sensitivity analysis [3], also known as the Sobol method, is one of the most widely used approaches for global sensitivity analysis of black-box models. The goal of the method is to estimate which proportion of the total variability in the model output can be attributed to one or more uncertain *inputs* (parameter values and/or initial conditions). Variance-based analysis requires the evaluation of multidimensional integrals, which in many cases cannot be calculated analytically. The numerical estimation of the integrals is commonly carried out using (Quasi-) Monte Carlo schemes [2]. These approaches demand a significant number of model evaluations to provide accurate approximations, especially in the case of complex, high-dimensional functions, and are therefore very computationally demanding.

T. Kurdyaeva and A. Milias-Argeitis are with the Groningen Biomolecular Sciences and Biotechnology Institute, University of Groningen, Nijenborgh 4, 9747 AG, Groningen, Netherlands. Emails: t.kurdyaeva@rug.nl, a.milias.argeitis@rug.nl

In this paper, we provide an alternative method for calculating sensitivity indices that relies on approximating the input-output relation of a given model with a Gaussian process (GP). GPs are a widely used tool for Bayesian nonlinear regression, as they are flexible, robust to overfitting and perform well using a relatively small number of model evaluations. For the purpose of variance-based sensitivity analysis, GPs offer an additional advantage, since they enable the *analytical* calculation of the sensitivity indices of interest, given a training set of model evaluations, and thus bypass the large number of evaluations required by Monte Carlo methods. This feature is particularly attractive when these evaluations are computationally intensive, as is, for example, the simulation of nonlinear biochemical network models.

The observation that GP regression can allow the analytical evaluation of variance-based sensitivity indices estimates was already made long ago [4], however both [4] and subsequent work [5] presented formulas for "generic" GP configurations and did not proceed all the way to the final results. In fact, a considerable amount of calculation is still necessary to arrive at these final formulas, as we show here for a specific and widely used GP configuration. In this paper, we thus present for the first time a complete analytic calculation of variance-based sensitivity indices estimates using GP metamodel. We then apply our findings to the study of steady-state global sensitivity analysis of a nonlinear biochemical reaction network and demonstrate the superiority of the GP-based approach over a state-of-the-art Quasi Monte Carlo-based software tool.

II. MATHEMATICAL BACKGROUND

A. Variance-based global sensitivity analysis

Consider a black-box model represented by a function $Y = F(\mathbf{X})$, where $\mathbf{X} := (X_1, \ldots, X_d)^T$ is a d-dimensional column vector of *input* variables and Y is the model *output*. In the following we will assume that the uncertainty in each $X_s, s = 1, ..., d$ can be modeled by a uniform distribution on the unit interval $[0, 1]$ and that the inputs are mutually independent. Assuming that F is square integrable with respect to the distribution of $\mathbf{X}$, we can then compute the sensitivity of Y to an individual input X_s through the the first-order or "main-effect" Sobol sensitivity index S_s, defined as

$$S_s = \frac{\mathrm{Var}\{E\left(Y|X_s\right)\}}{\mathrm{Var}\{Y\}},\ s = 1, \ldots, d, \tag{1}$$

where the expectation is taken with respect to all inputs except X_s and the variance term in the numerator is com-

puted with respect to the distribution of X_s. This quantity expresses the expected reduction in the variance of Y if we could learn and fix the value of X_s [3]. Therefore, S_s can help determine which uncertain inputs should be determined first to maximize the reduction in the variance of Y. In a completely analogous way, one can define sensitivity indices over groups of inputs. Further details can be found in [3]. It should be noted that each X_s may in general be assumed to follow a distribution other than the uniform. However, the independence assumption is necessary for an intuitive interpretation of the sensitivity results [2].

B. Gaussian process regression

In general terms, a GP defines a prior distribution over functions $f(\mathbf{x})$ which encapsulates modeling assumptions regarding smoothness, characteristic length scales, isotropy, periodicity etc. for a function F that we would like to learn. By observing the (possibly noise-corrupted) values of F at a set of *training points*, we can then calculate a posterior distribution that incorporates the information obtained from these observations. In mathematical terms, a GP is a collection of random variables indexed by a continuous variable $\mathbf{x} = (x_1, \ldots, x_d)^T$ (i.e. a continuous stochastic process) such that any finite number of these random variables has a multivariate Gaussian distribution. This definition implies that a GP is completely specified by its *mean* ($m(\mathbf{x})$) and *covariance* (or *kernel*) ($k(\mathbf{x}, \mathbf{x}')$) functions, defined as

$$m(\mathbf{x}) = \mathbb{E}\left[f(\mathbf{x})\right] \text{ and } k(\mathbf{x}, \mathbf{x}') = \text{Cov}\left[f(\mathbf{x}), f(\mathbf{x}')\right],$$

which allow us to write the GP as

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')).$$

While the mean is typically assumed to be zero, the choice of the kernel $k(\cdot, \cdot)$ is crucial, as it expresses our prior belief about the properties of F, the underlying function of interest. For example, if k is a smooth function, sample functions drawn from the GP will also be smooth, while the length scales over which k vanishes determines how noisy the sample functions will look [6].

Following the Bayesian approach, our prior beliefs are updated based on the information provided by the training data, denoted by $D = \{\widetilde{X}, \widetilde{\mathbf{y}}\}$ where $\widetilde{\mathbf{y}} = \left(\widetilde{y}^{(1)}, \ldots, \widetilde{y}^{(n)}\right)^T$ is a vector of function values $\widetilde{y}^{(i)} = F(\widetilde{\mathbf{x}}^{(i)})$ obtained at n points $\widetilde{X} = (\widetilde{\mathbf{x}}^{(1)}, \ldots, \widetilde{\mathbf{x}}^{(n)})$. To facilitate the presentation of the formulas below, for a set of points $X = \left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right)$, we will define $K(X, X)$ as the $n \times n$ matrix whose (i, j)-th element is equal to $k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)})$. In a similar fashion, $K(\mathbf{x}, X)$ is the $1 \times n$ vector with the i-th element being equal to $k(\mathbf{x}, \mathbf{x}^{(i)})$ and $f(X)$ is the vector of function values $f(\mathbf{x}^{(1)}), \ldots, f(\mathbf{x}^{(n)})$.

We can now present the closed form for the posterior predictive distribution of the GP at any set of *test* points $X_* = \left(\mathbf{x}_*^{(1)}, \ldots, \mathbf{x}_*^{(n_*)}\right)$ conditioned on the observed data D. The asterisk ($*$) will be used to denote both posterior quantities and the test points at which these are evaluated. Provided the likelihood of the data is also Gaussian, the posterior predictive distribution of function values f_* at X_* given the training data D is again a multivariate Gaussian:

$$f_*|D, X_* \sim \mathcal{N}(m(X_*), v(X_*)),$$

$$m(X_*) = K(X_*, \widetilde{X}) \left[K(\widetilde{X}, \widetilde{X}) + \sigma^2 I\right]^{-1} \widetilde{\mathbf{y}}, \tag{2}$$

$$v(X_*) = K(X_*, X_*) - \tag{3}$$

$$- K(X_*, \widetilde{X}) \left[K(\widetilde{X}, \widetilde{X}) + \sigma^2 I\right]^{-1} K(\widetilde{X}, X_*),$$

where the additive term $\sigma^2 I$ reflects the fact that the function values are corrupted by additive Gaussian noise with variance σ^2. Even if $\sigma^2 = 0$ (as is the case when the function is exactly evaluated), a small noise term is typically inserted to regularize the inversion of the covariance matrix $K(\widetilde{X}, \widetilde{X})$ [6]. The formula for the posterior predictive distribution mean and covariance will serve as a basis for the analytical computation of the variance-based first-order Sobol indices in the following section.

III. GP-BASED STEADY STATE SENSITIVITY ANALYSIS OF BIOCHEMICAL NETWORKS:SETUP

In this paper we consider a deterministic biochemical reaction network model described via a set of (nonlinear) ordinary differential equations (ODEs). The model contains r states, denoted by $\mathbf{z} = [z_1, ..., z_r]^T$, and p parameters, denoted by $\boldsymbol{\theta} = [\theta_1, ..., \theta_p]^T$. The temporal behavior of the system is then described by

$$\frac{d\mathbf{z}}{dt} = \phi(\mathbf{z}(t), \boldsymbol{\theta}). \tag{4}$$

In the following, we will assume that some of the system parameters and/or state initial conditions are not known with certainty, but are allowed to vary within some prespecified bounds. In line with the previously introduced terminology, we will collectively call these uncertain quantities uncertain *inputs* and denote them by a d-dimensional vector $\mathbf{X}$. We will further assume that the system has a unique equilibrium for each value of $\mathbf{X}$, which means that each component of the equilibrium vector $\mathbf{z}_{eq} = \lim_{t \to \infty} \mathbf{z}(t)$ is a function of $\mathbf{X}$. While in a restricted number of cases one may be able to analytically solve the algebraic equations to determine the equilibrium points, in most practically encountered cases this is not possible. We therefore shall assume that the values of $\mathbf{z}_{eq}(\mathbf{X})$ can only be obtained via forward simulation of (4) at a set of training points $\widetilde{X}$ (cf. with previous section). Hence, although the system equations are known, the map from $\mathbf{X}$ to $\mathbf{z}_{eq}$ is seen and treated as a black-box model.

To investigate how the system equilibrium is influenced by changes in $\mathbf{X}$, we will make use of variance-based global sensitivity analysis and calculate the first-order Sobol indices (defined in (1)) of $\mathbf{z}_{eq} = (z_{eq,1}, \ldots, z_{eq,r})^T$ with respect to the components of $\mathbf{X}$. To formally express our uncertainty about the inputs, we will assume that, after the necessary transformations (to be detailed in the examples below), the elements of the input vector $X_s, s = 1, ..., d$ are independent random variables which are uniformly distributed on the unit interval $[0, 1]$. The standard uniform distribution is

commonly used in global sensitivity analysis to avoid strict assumptions about input distributions when no additional information is present. In case more information on the inputs is available, other continuous distributions can be also assumed.

As we discussed in the previous section, the calculation of the Sobol indices will be based on Bayesian inference using GP regression. More specifically, we will consider a GP prior with zero mean and a pre-specified covariance function $k(\cdot,\cdot)$ for each state $z_{eq,1}(\mathbf{x}),\ldots,z_{eq,r}(\mathbf{x})$, where, to simplify the notation, Y represents any one of the equilibrium state components. For the covariance function, we will make use of the popular *squared exponential* (SE) kernel:

$$k(\mathbf{x},\mathbf{x}') = \sigma_f^2 \prod_{j=1}^{d} \exp\left(-\frac{(x_j - x_j')^2}{2\,l_j^2}\right) \tag{5}$$

This function is characterized by two *hyperparameters*: $\mathbf{l} = (l_1,...,l_d)$ is a vector of *length scales* and controls the scales over which the sample functions are allowed to fluctuate along the different coordinates. On the other hand, the *scale factor* σ_f denotes the maximum allowable covariance of the sample functions, and thus determines the range of variation of sample functions.

Since the hyperparameters have a large effect on the form of functions that a GP can generate, proper choice of the hyperparameters is critical for avoiding overfitting the regression data and favoring more parsimonious GP models. To choose hyperparameter values for a given training dataset D, a common strategy is to maximize the *marginal likelihood* of the function values $\widetilde{\mathbf{y}}$ conditioned on the hyperparameters and the training points $\widetilde{X}$; that is, maximize $p(\widetilde{\mathbf{y}}|\widetilde{X},\theta)$, where θ contains all the tunable hyperparameters of the kernel.

To approximate the unknown function Y, we can therefore generate a training dataset $D = \{\widetilde{X},\widetilde{\mathbf{y}}\}$ consisting of the equilibrium states at n input points $\widetilde{X} = (\widetilde{\mathbf{x}}^{(1)},\ldots,\widetilde{\mathbf{x}}^{(n)})$. This information will be used to arrive at the mean and variance of the posterior process ((2) and (3)) by optimizing over the kernel hyperparameters. Our formulas for the sensitivity indices will then be based on these posterior mean and variance formulas.

Besides its simple interpretability, the common choice of the SE kernel is motivated by a number of useful properties it possesses, such as the fact that it is a so-called *universal* kernel [6], capable of uniformly approximating any continuous function under certain conditions. Furthermore, the form of the SE kernel together with the assumption of independent, uniformly distributed inputs allow us to perform the calculation of the first-order Sobol indices analytically, as we will show next.

IV. GP-BASED STEADY STATE SENSITIVITY ANALYSIS OF BIOCHEMICAL NETWORKS: CALCULATIONS

In this section we outline the main steps for obtaining the closed form estimates the first-order Sobol indices $S_s, s = 1,...,d$ defined in (1) using GP regression.

As described previously, our starting point is the posterior GP approximation of the unknown function Y which connects the model inputs to the output of interest. This posterior distribution, obtained by evaluating the model at a set of training points and optimizing the kernel hyperparameters (cf. with Section III), is described by the mean and variance functions given in (2) and (3), respectively. Note that, since we are approximating Y by a stochastic process, the variances appearing in (1) are random quantities. Here, we are interested in the posterior mean of these quantities (given the training data set), although their variance is in principle analytically computable as well. The posterior GP mean in (2), which becomes a deterministic function given the set of training input points, will serve as a basis for calculating the posterior mean of $\text{Var}\{E(Y|X_s)\}$ for $s = 1,\ldots,d$, and $\text{Var}\{Y\}$ given the training data. Denoting all expectations with respect to the GP posterior distribution by an asterisk $(*)$, the GP-based Sobol index estimates can then be calculated as

$$\hat{S}_s = \frac{\mathbb{E}_*\{\text{Var}\{E(Y|X_s)\}\}}{\mathbb{E}_*\{\text{Var}(Y)\}},\ s = 1,...,d. \tag{6}$$

The analytic derivation of the closed-form expressions for these estimates is too lengthy to be presented here and can be found in [7]. Finally, we should point out that our calculations are made possible thanks to the closed-form expressions for the GP posterior mean and variance given the set of training input points and the optimized kernel hyperparameters.

V. EXAMPLES

We will present two examples to demonstrate the accuracy and efficiency of our approach to global sensitivity analysis. First, we will consider a commonly used benchmark function with known Sobol indices. Then, we will calculate the first-order Sobol indices for a mathematical model of a bacterial two-component signaling system and compare the accuracy and efficiency of our approach with that of a recently presented Quasi Monte Carlo-based method.

A. The Sobol g-function

The Sobol g-function is commonly used as benchmark in variance-based sensitivity studies [5]. For a d-dimensional input vector $(X_1,...,X_d) \in [0,1]^d$, it is defined as:

$$Y = g(X_1,...,X_d) = \prod_{j=1}^{d} g_j(X_j),\ g_j(X_j) = \frac{|4X_j - 2| + a_j}{1 + a_j}, \tag{7}$$

where each $a_j \geq 0$ controls the relative importance of the input X_j: the larger a_j is, the less X_j affects the value of the function. Even though the g-function is nonlinear and non-monotonic, the theoretical values of Sobol indices are easy to compute and can thus be used to benchmark algorithms for variance-based sensitivity analysis.

To test our method, we will consider a 4-dimensional g-function with coefficients $a_1 = 0, a_2 = 1, a_3 = 2, a_4 = 4$,

and assume that the inputs are mutually independent and uniformly distributed in $[0, 1]$. According to these assumptions, X_1 contributes the most to the variance of Y, X_2 and X_3 are less significant, while X_4 is almost insignificant. This is reflected in the theoretical values of the first-order Sobol indices, provided in Table I. Analytical formulas for calculating the indices can be found in [2].

The GP regression was implemented in the Matlab GPML toolbox [8]. The GP was trained with a varying number of data points by evaluating the g-function on a point set obtained from the Sobol low-discrepancy sequence [9] (implemented via the `sobolset` Matlab function). For every choice of training sample size, the first-order Sobol indices were calculated and then averaged over 100 independent runs. Table I contains the mean values (and standard deviation in brackets) of the first-order Sobol indices with respect to each input $X_i, i = 1, ..., 4$, together with their true values.

TABLE I

GP-BASED SOBOL INDEX ESTIMATES FOR THE g-FUNCTION

	First-order Sobol indices $\hat{S}$				**Sample size**
Input	X_1	X_2	X_3	X_4	
Estimate	0.7574 (0.2019)	0.0609 (0.0982)	0.0233 (0.0598)	0.0146 (0.0210)	25
	0.6514 (0.0784)	0.1484 (0.0518)	0.0404 (0.0322)	0.0110 (0.0182)	50
	0.6558 (0.0223)	0.1497 (0.0182)	0.0632 (0.0125)	0.0068 (0.0572)	100
	0.6464 (0.0096)	0.1571 (0.0084)	0.0701 (0.0043)	0.0197 (0.0029)	200
Exact	**0.6436**	**0.1609**	**0.0715**	**0.0257**	

According to the results in Table I, increasing the training sample size makes the estimated index values approach the true ones with less variability. Remarkably, a training data set of only 200 points already yields highly accurate estimates. For smaller numbers of training points, even though the predictions deviate from the theoretical values, the ordering of the true sensitivity indices (reflecting the importance of the corresponding inputs) is identified correctly.

B. A bacterial two-component system model

Here, we will test our GP-based global sensitivity method on a nonlinear ODE model of a bacterial two-component system (TCS). TCSs serve as a ubiquitous signaling mechanism with which bacterial cells regulate their behavior with respect to physcial or chemical changes in their environment. They consist of a membrane-bound histidine kinase (H) that senses specific environmental stimulus, and its cognate response regulator (R) that mediates the cellular response, typically by controlling the expression of several target genes. Upon detection of a particular stimulus, H gets autophosphorylated (H_p) and transfers the phosphoryl group to the response regulator (R_p), causing its activation. In many cases, the unphosphorylated form of the histidine kinase (H) also acts as phosphatase, which dephosphorylates (and thus deactivates) R_p. The basic reaction scheme corresponding to these biochemical processes, taken from [10], is presented in Fig. 1.

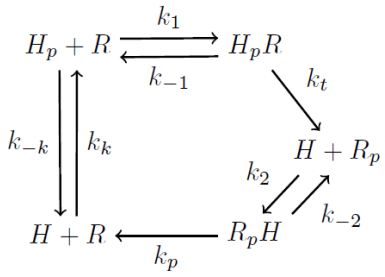

Fig. 1. Schematic illustration of TCS kinetics

A key property of TCSs, first predicted mathematically [10] and verified experimentally [10], [11], is that the steady-state level of R_p is largely insensitive to variations in the concentrations of the histidine kinase and the response regulator when the response regulator concentration is sufficiently large compared to the kinase (e.g. more than 10-fold).

The mathematical model describing the concentration dynamics of a TCS consists of a system with four states, H_p, R_p, and the concentrations of the complexes, H_pR and R_pH. Since the total amounts of kinase and response regulator (H_t and R_t) are assumed to remain constant over time, the four ODEs are accompanied by two algebraic constraints, as presented in [10].

To further illustrate the utility of our GP-based steady-state sensitivity approach, we will investigate the robustness of the steady-state R_p levels with respect to a number of parameters and initial conditions in the model. Our method will be compared in terms of accuracy and efficiency with a commonly used algorithm for Sobol index estimation, based on Quasi-Monte Carlo.

For our simulations we will consider a set of biologically plausible nominal parameter values and initial conditions provided in [12]. Given that parameter values for TCS models are estimated experimentally with a large degree of uncertainty, it is not uncommon for the estimates to vary by an order of magnitude or more. To capture this phenomenon, we assumed that whenever the TCS model parameters and initial conditions are uncertain, they follow a log-uniform distribution defined as follows: given the nominal value of an input, x'_0, we consider the log-ratio $x := \log_{10}(x'/x'_0)$, and assume that $x \sim \mathcal{U}([-0.5, 0.5])$, meaning that the log-ratio varies over an order of magnitude. This results in a log-uniform distribution for $x' = x'_0 \cdot 10^x$. From now on, when we refer to model inputs, we imply the log-ratios x.

In our first test, we assumed that four important inputs to the TCS model are uncertain. These are the two initial conditions, H_t and R_t, and two parameters, the phosphotransfer rate (k_t) and the dephosphorylation rate of the histidine kinase (k_p). We were interested in the effects of these uncertain inputs on the steady-state concentration of R_p. Using the Matlab IQM toolbox [13] which provides fast simulation of nonlinear ODEs, we thus simulated the mathematical model

of TCS at sets of input samples with varying sizes, obtained from a Sobol low-discrepancy sequence. For every training set we then built a GP model using the Matlab GPML toolbox. Finally, we calculated the first-order Sobol indices of R_p with respect to each uncertain input.

Since analytical expressions for the Sobol indices cannot be computed in this case, to access the accuracy and efficiency of our method we calculated the indices ten times for each input sample size (by randomly scrambling the Sobol sequence) and computed the corresponding means and standard deviations. We then compared our results with a recently presented Matlab implementation of the Quasi-Monte Carlo algorithm for Sobol index calculation, GSAT [14] (as in the GP case, all necessary model evaluations were performed with the IQM toolbox). Given that the bias of Quasi-Monte Carlo estimates decreases considerably with the number of samples, GSAT should provide unbiased estimates of the "true" Sobol index values for very large sample sizes. As in the case of GP-based sensitivity, we ran GSAT for sample sets of varying size, repeating the calculation ten times for each sample size. Since Monte Carlo-based estimates of Sobol indices are grossly inaccurate for small sample sizes, the GSAT sample sizes were ∼10-fold larger compared to the GP algorithm. The results from the comparison of the two methods are summarized in Table II, which displays the mean estimates together with the corresponding standard deviations in parentheses.

TABLE II

FIRST-ORDER SENSITIVITY INDICES OF R_p IN THE TCS MODEL

	First-order Sobol indices $\hat{S}$				Sample size	Average time (s)
Input	k_t	k_p	H_t	R_t		
GP method	0.0026 (0.0063)	0.8630 (0.0556)	0.0104 (0.0056)	0.0292 (0.0185)	25	0.30
	0.0000 (0.0000)	0.8313 (0.0132)	0.0076 (0.0017)	0.0470 (0.0030)	100	0.71
	0.0000 (0.0000)	0.8306 (0.0022)	0.0078 (0.0002)	0.0468 (0.0005)	400	6.50
GSAT	-0.0027 (0.0064)	0.8257 (0.0167)	0.0058 (0.0067)	0.0462 (0.0046)	1000	5.59
	0.0013 (0.0021)	0.8315 (0.0023)	0.0094 (0.0021)	0.0476 (0.0015)	4000	23.55
	-0.0005 (0.0008)	0.8315 (0.0006)	0.0077 (0.0004)	0.0460 (0.0006)	16000	95.01

Based on the results of Table II, both our GP-based method and GSAT clearly show that k_p is the most influential factor on the steady-state levels of R_p. Moreover, the concentration of R_p is quite robust to changes in H_t and R_t, as indicated by the small values of the corresponding Sobol indices. This observation is in agreement with the main findings of [10]. Finally, the effect of k_t on R_p is insignificant in comparison to other input parameters.

By comparing the sample sizes used by the two methods, we can see that our GP-based approach requires considerably less model evaluations to obtain approximately the same accuracy as GSAT. In particular, the GP needs around 400 model evaluations to produce estimates as precise as those provided by GSAT with 16000 samples. This results in a significant reduction in computational cost, since the average running time drops from 95 seconds with GSAT to just 6.5 seconds using our approach. Finally, we should point out that Monte Carlo-based estimates of Sobol indices may turn out negative in case the actual index value is very small. This is a well-documented behavior [2].

As a final test of the relative efficiency of our method, we compared the accuracy of the Sobol index estimates when allowing both methods to run for the *same* amount of time. Besides the previous setup with four uncertain inputs, we also increased the complexity of the problem by assuming that *all* eight model parameters and the two initial conditions are uncertain. For the scenario with four uncertain inputs, we set an average running time of 6.5 seconds for both methods. Within this time, the GP-based algorithm uses 400 training points while GSAT makes 1400 model evaluations. In the case of ten uncertain inputs, both algorithms were run for 69.5 seconds on average. In this case, 1000 and 6400 data points were used by the GP method and GSAT respectively.

For each estimation scenario, we performed ten independent runs of each method and calculated the corresponding statistics for the first-order Sobol indices of R_p with respect to H_t and R_t. The distributions of the obtained estimates are summarized with boxplots in Fig. 2. As can be readily observed, our GP-based method yields noticeably more accurate estimates.

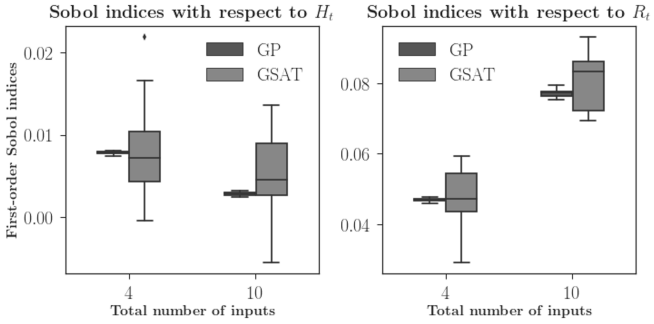

Fig. 2. Boxplots of first-order Sobol index estimates of R_p with respect to H_t and R_t, obtained from ten runs of each method

To further investigate the case with ten uncertain inputs (eight model constants and two initial conditions), we calculated the GP-based first-order Sobol indices for all four states. Fig.3 represents graphically the estimates of sensitivity measures from one run using 1000 training data points. The color intensity corresponds to the effect size of every input on the corresponding state where the darker shades of grey indicate the higher impact. For instance, as we can see, the concentration of R_pH is very sensitive to the levels of H_t and quite robust to R_t.

VI. DISCUSSION

Variance-based sensitivity indices are a powerful, generally applicable and intuitively interpretable approach for

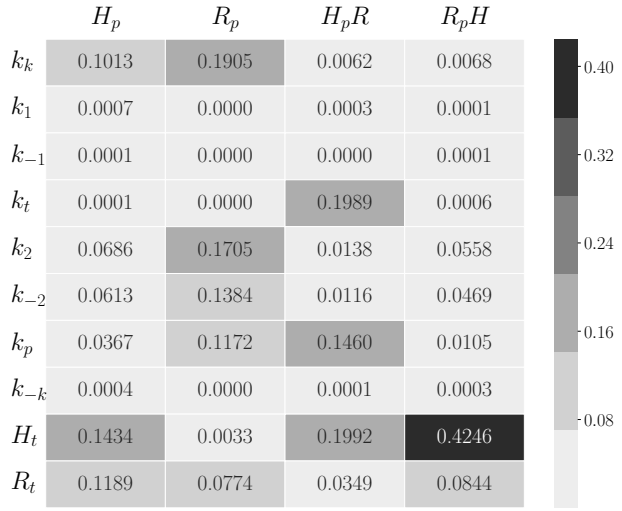

Fig. 3. Heatmap of first-order Sobol index estimates obtained from GP-based approach for four states (H_p, R_p, H_pR, R_pH) with respect to all ten uncertain inputs (eight model parameters and two initial conditions)

global sensitivity analysis. Their main drawback is computational cost, as they require thousands of model evaluations to obtain accurate estimates. We presented an alternative approach to variance-based analysis using GP regression that offers a much more efficient alternative to commonly used Monte Carlo methods. Our method requires the evaluation of a model at a set of training inputs. This data can in turn be used to infer a GP posterior distribution that approximates the input-output behavior of the model. Finally, as we have shown, the integrals required for the estimation of the Sobol indices can be calculated analytically, thus bypassing the need for Monte Carlo integration. The main limitation of GPs is that they require the repeated inversion of large matrices during the training step, but there is already a large body of literature addressing this task (e.g. [15]).

Another approach for global sensitivity analysis using GP models was introduced in [16]. In this paper, the extensive computational costs of evaluating a computationally expensive models were significantly reduced by first constructing a GP-based approximation and using it to compute the Sobol indices via Monte Carlo. Our method allows us to reduce the number of model runs even further by using the exact analytical formulas for the Sobol indices rather than MC sampling schemes.

Alternatively, Sobol indices can be appproximated with *polynomial chaos expansion* (PCE). However, PCE is prone to overfitting when high-degree polynomials are included in the expansion and cannot estimate the prediction uncertainty in the output [17]. On the other hand, the GP-based approach gives a natural trade-off between the fit to the data and the model complexity to prevent from overfitting. Additionally, it provides a full probabilistic description of the uncertain output.

Analytical calculations of global sensitivity measures may be also performed using the *unscented transform* (UT). Even though the UT was shown to be more efficient than MC techniques in many applications, its accuracy decreases drastically for more complex problems with large number of uncertain inputs [18].

Even though we presented analytic formulas for computing the first-order Sobol indices, our formulas are immediately generalizable to high-order indices [2] by considering groups of inputs instead of a single input at a time. In this way, the effects of varying several input simultaneously can be investigated. Our approach can be especially useful for studying the global sensitivity of computationally expensive high-dimensional biochemical models, not only at steady-state, but at any desired time point.

REFERENCES

[1] Z. Zi, "Sensitivity analysis approaches applied to systems biology models," *IET Systems Biology*, vol. 5, no. 6, pp. 336–346, 2011.

[2] A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, and S. Tarantola, *Global sensitivity analysis: the primer*. John Wiley & Sons, 2008.

[3] I. M. Sobol, "Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates," *Mathematics and computers in simulation*, vol. 55, no. 1-3, pp. 271–280, 2001.

[4] J. E. Oakley and A. O'Hagan, "Probabilistic sensitivity analysis of complex models: a Bayesian approach," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 66, no. 3, pp. 751–769, 2004.

[5] A. Marrel, B. Iooss, B. Laurent, and O. Roustant, "Calculations of Sobol indices for the Gaussian process metamodel," *Reliability Engineering & System Safety*, vol. 94, no. 3, pp. 742–751, 2009.

[6] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[7] [Online]. Available: http://hdl.handle.net/11370/2aff37c2-821e-4bfd-aaf5-33296e419840

[8] C. E. Rasmussen and H. Nickisch, "Gaussian processes for machine learning (GPML) toolbox," *Journal of Machine Learning Research*, vol. 11, pp. 3011–3015, 2010.

[9] I. Sobol, "Quasi-Monte Carlo methods," *Progress in Nuclear Energy*, vol. 24, no. 1, pp. 55 – 61, 1990.

[10] E. Batchelor and M. Goulian, "Robustness and the cycle of phosphorylation and dephosphorylation in a two-component regulatory system," *Proceedings of the National Academy of Sciences*, vol. 100, no. 2, pp. 691–696, 2003.

[11] R. Gao and A. M. Stock, "Probing kinase and phosphatase activities of two-component systems in vivo with concentration-dependent phosphorylation profiling," *Proceedings of the National Academy of Sciences*, vol. 110, no. 2, pp. 672–677, 2013.

[12] O. A. Igoshin, R. Alves, and M. A. Savageau, "Hysteretic and graded responses in bacterial two-component signal transduction," *Molecular microbiology*, vol. 68, no. 5, pp. 1196–1215, 2008.

[13] IQM Tools. [Online]. Available: http://www.intiquan.com/iqm-tools/

[14] F. Cannavó, "Sensitivity analysis for volcanic source modeling quality assessment and model selection," *Computers & Geosciences*, vol. 44, pp. 52–59, 2012.

[15] M. Titsias, "Variational learning of inducing variables in sparse Gaussian processes," in *Artificial Intelligence and Statistics*, 2009, pp. 567–574.

[16] J. Rohmer and E. Foerster, "Global sensitivity analysis of large-scale numerical landslide models based on Gaussian-Process meta-modeling," *Computers & geosciences*, vol. 37, no. 7, pp. 917–927, 2011.

[17] A. O'Hagan, "Polynomial chaos: A tutorial and critique from a statisticians perspective," *SIAM/ASA J. Uncertainty Quantification*, vol. 20, pp. 1–20, 2013.

[18] X. Cheng and V. Monebhurrun, "Application of different methods to quantify uncertainty in specific absorption rate calculation using a CAD-based mobile phone model," *IEEE Transactions on Electromagnetic Compatibility*, vol. 59, no. 1, pp. 14–23, 2017.