Journal of Theoretical Biology 254 (2008) 178–196

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Review

A methodology for performing global uncertainty and sensitivity analysis in systems biology

Simeone Marino, Ian B. Hogue, Christian J. Ray, Denise E. Kirschner*

Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, MI 48109-0620, USA

ARTICLE INFO

Article history:
Received 6 December 2007
Received in revised form
14 March 2008
Accepted 12 April 2008
Available online 20 April 2008

Keywords:
Latin hypercube sampling (LHS)
Partial rank correlation coefficient (PRCC)
Extended Fourier amplitude sensitivity test (eFAST)
Agent-based model (ABM)
Sensitivity index
Monte Carlo methods
Aleatory uncertainty
Epistemic uncertainty

ABSTRACT

Accuracy of results from mathematical and computer models of biological systems is often complicated by the presence of uncertainties in experimental data that are used to estimate parameter values. Current mathematical modeling approaches typically use either single-parameter or local sensitivity analyses. However, these methods do not accurately assess uncertainty and sensitivity in the system as, by default, they hold all other parameters fixed at baseline values. Using techniques described within we demonstrate how a multi-dimensional parameter space can be studied globally so all uncertainties can be identified. Further, uncertainty and sensitivity analysis techniques can help to identify and ultimately control uncertainties. In this work we develop methods for applying existing analytical tools to perform analyses on a variety of mathematical and computer models. We compare two specific types of global sensitivity analysis indexes that have proven to be among the most robust and efficient. Through familiar and new examples of mathematical and computer models, we provide a complete methodology for performing these analyses, in both deterministic and stochastic settings, and propose novel techniques to handle problems encountered during these types of analyses.

Contents

* Corresponding author. Tel.: +1734 647 7722; fax: +1734 647 7723.
E-mail address: kirschne@umich.edu (D.E. Kirschner).

0022-5193/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.jtbi.2008.04.011

1. Introduction

Systems biology is the study of the interactions between the components of a biological system, and how these interactions give rise to the function and behavior of the system as a whole. The systems biology approach often involves the development of mathematical or computer models, based on reconstruction of a dynamic biological system from the quantitative properties of its elementary building blocks. Building mathematical and computational models is necessary to help decipher the massive amount of data experimentalists are uncovering today. The goal of the systems biologist or modeler is to represent, abstract, and ultimately understand the biological world using these mathematical and computational tools. Experimental data that are available for each system should guide, support, and shape the model building process. This can be a daunting task, especially when the components of a system form a very complex and intricate network.

Paraphrasing Albert Einstein, models should be as simple as possible, but not simpler. A parsimonious approach must be followed. Otherwise, if every mechanism and interaction is included, the resulting mathematical model will be comprised of a large number of variables, parameters, and constraints, most of them uncertain because they are difficult to measure experimentally, or are even completely unknown in many cases. Even when a parsimonious approach is followed during model building, available knowledge of phenomena is often incomplete, and experimental measures are lacking, ambiguous, or contradictory. So the question of how to address uncertainties naturally arises as part of the process. Uncertainty and sensitivity (US) analysis techniques help to assess and control these uncertainties.

Uncertainty analysis (UA) is performed to investigate the uncertainty in the model output that is generated from uncertainty in parameter inputs. Sensitivity analysis (SA) naturally follows UA as it assesses how variations in model outputs can be apportioned, qualitatively or quantitatively, to different input sources (Saltelli et al., 2000). In this work we review US analysis techniques in the context of deterministic dynamical models in biology, and propose a novel procedure to deal with a particular stochastic, discrete type of dynamical model (i.e. an agent-based model—ABM[1]).

By deterministic model, we mean that the output of the model is completely determined by the input parameters and structure of the model. The same input will produce the same output if the model were simulated multiple times. Therefore, the only uncertainty affecting the output is generated by input variation. This type of uncertainty is termed *epistemic* (or subjective, reducible, type B uncertainty; see Helton et al., 2006). Epistemic uncertainty derives from a lack of knowledge about the adequate value for a parameter/input/quantity that is assumed to be constant throughout model analysis. In contrast, a stochastic model will not produce the same output when repeated with the same inputs because of inherent randomness in the behavior of the system. This type of uncertainty is termed *aleatory* (or stochastic, irreducible, type A; see Helton et al., 2006). This distinction has been and still is an area of interest and study in the engineering and risk assessment community (see Apostolakis, 1990; Helton, 1997; Helton et al., 2007; Parry and Winter, 1981; Pate'-Cornell, 1996).

Many techniques have been developed to address US analysis: differential analysis, response surface methodology, Monte Carlo (MC) analysis, and variance decomposition methods. See Iman and Helton, (1988) and Saltelli et al. (2000) for details on each of these approaches and Cacuci and Ionescu-Bujor (2004), Draper (1995), Helton (1993) and Saltelli et al. (2005) for more general reviews on US analysis. Here we briefly illustrate the most popular, reliable, and efficient UA techniques and SA indexes. In Section 2, we describe two UA techniques: a MC approach and Latin hypercube sampling (LHS). In Section 3, we describe two SA indexes: partial rank correlation coefficient (PRCC) and extended Fourier amplitude sensitivity test (eFAST): PRCC is a sampling-based method, while eFAST is a variance-based method. In Section 4, we perform US analysis on both new and familiar deterministic dynamical models (quantifying epistemic uncertainty) from epidemiology and immunology, and discuss results. Section 5 presents an ABM, where we suggest a method to deal with the aleatory uncertainty that results from the stochasticity embedded in the model structure, to facilitate the use of PRCC and eFAST techniques. We use Matlab (Copyright 1984–2006 The MathWorks, Inc., Version 7.3.0.298 R2006b) to solve all the differential equation systems of Section 4 and to implement most of the US analysis functions described throughout the manuscript (available on our website, http://malthus.micro.med.umich.edu/lab/usanalysis.html).

2. Uncertainty analysis

Input factors for most mathematical models consist of parameters and initial conditions for independent and dependent model variables. As mentioned, these are not always known with a sufficient degree of certainty because of natural variation, error in measurements, or simply a lack of current techniques to measure them. The purpose of UA is to quantify the degree of confidence in the existing experimental data and parameter estimates. In this section we describe the most popular sampling-based approaches used to perform UA, MC methods, and their most efficient implementation, namely the LHS technique.

2.1. Monte Carlo simulation

MC methods are popular algorithms for solving various kinds of computational problems. They include any technique of statistical sampling employed to approximate solutions to quantitative problems. A MC simulation is based on performing multiple model evaluations using random or pseudo-random numbers to sample from probability distributions of model inputs. The results of these evaluations can be used to both determine the uncertainty in model output and perform SA. A large body of literature exists on the use of expert review processes to characterize epistemic uncertainty associated with poorly known model parameters (see for example Cooke, 1991; Evans et al., 1994; Hora and Iman, 1989; McKay and Meyer, 2000).

[1] IBM = Individual Based Modeling in fields like ecology.

For each parameter, sampling is guided by the specification of a probability density function (pdf) (i.e. normal, uniform, lognormal, etc.), depending on *a priori* information. If there are no *a priori* data, a natural choice is a uniform distribution (assigning some hypothetical, but large range with minimum and maximum values for the parameters). If biological knowledge exists suggesting a more frequent or expected value for a parameter, a normal pdf would be the best choice (setting the variance of the distribution as large as needed).

Several sampling strategies can be implemented to perform UA, such as random sampling, importance sampling, or LHS (Helton and Davis, 2003; Mckay et al., 1979). To recreate input factor distributions through sampling, a large number of samples are likely required. If too few iterations are performed, not all values may be represented in the samples or values in the outer ranges may be under-sampled. The LHS algorithm was specifically developed to address this problem and it is by far the most popular sampling scheme for UA (Morris, 2000).

2.2. *Latin hypercube sampling—LHS*

LHS belongs to the MC class of sampling methods, and was introduced by Mckay et al. (1979). LHS allows an un-biased estimate of the average model output, with the advantage that it requires fewer samples than simple random sampling to achieve the same accuracy (Mckay et al., 1979). LHS is a so-called *stratified sampling without replacement* technique, where the random parameter distributions are divided into *N* equal probability intervals, which are then sampled. *N* represents the sample size. The choice for *N* should be at least *k*+1, where *k* is the number of parameters varied, but usually much larger to ensure accuracy. If the interval of variation for some parameter is very large (several orders of magnitude), the sampling can be performed on a log scale to prevent under-sampling in the outer ranges of the interval where the parameter assumes very small values (see Supplement C and Fig. C.1 and C.2 online).

The LHS method assumes that the sampling is performed independently for each parameter, although a procedure to impose correlations on sampled values has also been developed (Iman and Conover, 1982; Iman and Davenport, 1982). The sampling is done by randomly selecting values from each pdf (Fig. 1A). Each interval for each parameter is sampled exactly once (without replacement), so that the entire range for each parameter is explored (Fig. 1A). A matrix is generated (which we call the LHS matrix) that consists of *N* rows for the number of simulations (sample size) and of *k* columns corresponding to the number of varied parameters (Fig. 1B). *N* model solutions are then simulated, using each combination of parameter values (each row of the LHS matrix, Fig. 1B).

The model output of interest is collected for each model simulation. Different model outputs can be studied if more than one model output is of interest.

3. Sensitivity analysis

SA is a method for quantifying uncertainty in any type of complex model. The objective of SA is to identify critical inputs (parameters and initial conditions) of a model and quantifying how input uncertainty impacts model outcome(s). When input factors such as parameters or initial conditions are known with little uncertainty, we can examine the partial derivative of the output function with respect to the input factors. This sensitivity measure can easily be computed numerically by performing multiple simulations varying input factors around a nominal value. This technique is called a local SA because it investigates

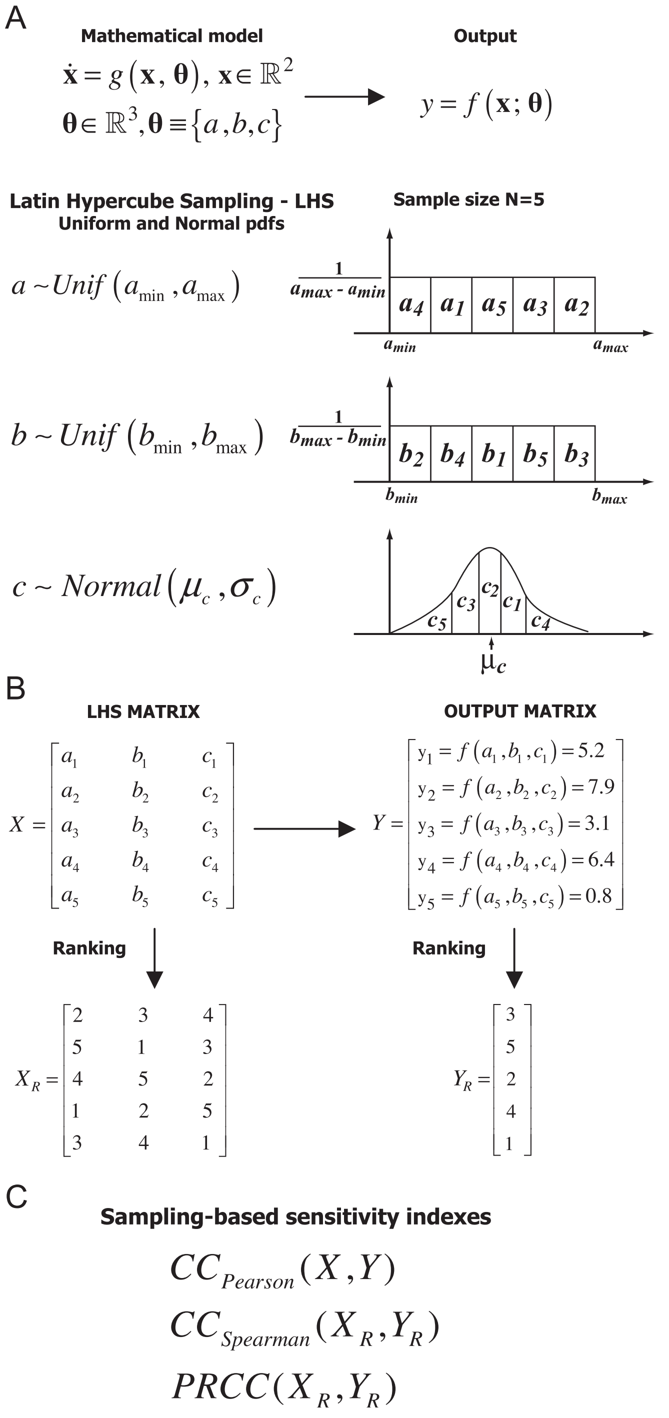
A
Mathematical model
$\dot{\mathbf{x}} = g(\mathbf{x}, \boldsymbol{\theta}),\ \mathbf{x} \in \mathbb{R}^2$
$\boldsymbol{\theta} \in \mathbb{R}^3, \boldsymbol{\theta} \equiv \{a, b, c\}$
Output
$y = f(\mathbf{x}; \boldsymbol{\theta})$
Latin Hypercube Sampling - LHS
Uniform and Normal pdfs
Sample size N=5
$a \sim Unif(a_{\min}, a_{\max})$
$\frac{1}{a_{max} - a_{min}}$
a_4 a_1 a_5 a_3 a_2
a_{min}
a_{max}
$b \sim Unif(b_{\min}, b_{\max})$
$\frac{1}{b_{max} - b_{min}}$
b_2 b_4 b_1 b_5 b_3
b_{min}
b_{max}
$c \sim Normal(\mu_c, \sigma_c)$
c_5 c_3 c_2 c_1 c_4
μ_c
B
LHS MATRIX
$X = \begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \\ a_4 & b_4 & c_4 \\ a_5 & b_5 & c_5 \end{bmatrix}$
OUTPUT MATRIX
$Y = \begin{bmatrix} y_1 = f(a_1, b_1, c_1) = 5.2 \\ y_2 = f(a_2, b_2, c_2) = 7.9 \\ y_3 = f(a_3, b_3, c_3) = 3.1 \\ y_4 = f(a_4, b_4, c_4) = 6.4 \\ y_5 = f(a_5, b_5, c_5) = 0.8 \end{bmatrix}$
Ranking
Ranking
$X_R = \begin{bmatrix} 2 & 3 & 4 \\ 5 & 1 & 3 \\ 4 & 5 & 2 \\ 1 & 2 & 5 \\ 3 & 4 & 1 \end{bmatrix}$
$Y_R = \begin{bmatrix} 3 \\ 5 \\ 2 \\ 4 \\ 1 \end{bmatrix}$
C
Sampling-based sensitivity indexes
$CC_{Pearson}(X, Y)$
$CC_{Spearman}(X_R, Y_R)$
$PRCC(X_R, Y_R)$

Fig. 1. Scheme of uncertainty and sensitivity analysis performed with LHS and PRCC methods. The mathematical model is represented as an ordinary differential equation system, where **x** is the vector of state variables in an *n*-dimensional space $\mathbb{R}^n$. (as an example we set $n = 2$ and $\boldsymbol{\theta}$ is the parameter vector in $\mathbb{R}^k$ ($k = 3$ in this example). For ease of notation, the output *y* is unidimensional and it is a function of **x** and $\boldsymbol{\theta}$. (A) Mathematical model specification (dynamical system, parameters, output) and the corresponding LHS scheme. Probability density functions (pdfs) are assigned to the parameters of the model (e.g. *a*, *b*, *c*). We show an example with sample size *N* equal to 5. Each interval is divided into five equiprobable subintervals, and independent samples are drawn from each pdf (uniform and normal). The subscript represents the sampling sequence. (B) The LHS matrix (*X*) is then built by assembling the samples from each pdf. Each row of the LHS matrix represents a unique combination of parameter values sampled without replacement. The hypothetical model $\dot{\mathbf{x}} = g(\mathbf{x}, \boldsymbol{\theta})$ is then solved, the corresponding output generated, and stored in the matrix *Y*. Each matrix is then rank-transformed (X_R and Y_R). (C) The LHS matrix (*X*) and the output matrix (*Y*) are used to calculate the Pearson correlation coefficient ($CC_{Pearson}$). The rank-transformed LHS matrix (X_R) and output matrix (Y_R) are used to calculate the Spearman or rank correlation coefficient (RCC) and the partial rank correlation coefficient (PRCC) (see Section 3.1).

the impact on model output, based on changes in factors only very close to the nominal values. In biology, input factors are often very uncertain and therefore local SA techniques are not appropriate for a quantitative analysis; instead global SA techniques are needed. These global techniques are usually implemented using MC simulations and are, therefore, called sampling-based methods.

Different SA techniques will perform better for specific types of mathematical and computational models. A natural starting point in the analysis with sampling-based methods would be to examine scatter plots. Scatter plots enable graphic detection of nonlinearities, non-monotonicities, and correlations between model inputs and outputs. See Helton and Davis (2002), Helton et al. (2006), Hora and Helton (2003), Kleijnen and Helton (1999), and Storlie and Helton (2008a, b) for a review on the indexes listed in the next paragraph.

For linear trends, linear relationship measures that work well are the Pearson correlation coefficient (CC), partial correlation coefficients (PCCs), and standardized regression coefficients (SRC). For nonlinear but monotonic relationships between outputs and inputs, measures that work well are based on rank transforms[2] such as Spearman rank correlation coefficient (RCC or Spearman's rho), partial rank correlation coefficient (PRCC), and standardized rank regression coefficients (SRRC). For nonlinear non-monotonic trends, methods based on decomposition of model output variance are the best choice. Examples of these methods are the Sobol method and its extended version based on (quasi) random numbers and an *ad hoc* design (see Saltelli, 2002 for details), the Fourier amplitude sensitivity test (FAST) and its extended version (eFAST). Aside from those listed, there are alternative methods available that are less affected by non-monotonic relationships between the inputs and the output, e.g. common means, common distributions or locations, common medians, and statistical independence. These methods are based on gridding (placing grids on a scatter plot) to evaluate any non-randomness in the distribution of points across the grid cells and they are generally less computationally expensive than variance-based methods (such as eFAST).

In general, the computational execution time of the model is the major concern when performing US analysis. Screening methods, such as those of Morris (1991), are global and computationally compatible: they represent adequate available tools to efficiently address the problem, if the model is very large and the execution time is prohibitive (several hours or days), as it is usually the case for ABMs (see below).

We will focus and implement only PRCC and eFAST as two examples of SA methods. PRCC and SRRC appear to be, in general, the most efficient and reliable (giving similar results) among the sampling-based indexes (see Saltelli and Marivoet, 1990) while eFAST has proven to be one of the most reliable methods among the variance-based techniques (Saltelli, 2004), although computationally expensive (see Ratto et al., 2007; Tarantola et al., 2006).

It is important to note that PRCCs and variance decompositions obtained with eFAST measure two very different model properties. Specifically, PRCCs provide a measure of monotonicity after the removal of the linear effects of all but one variable. In contrast, the results obtained with eFAST return measures of fractional variance accounted for by individual variables and groups of variables. Ideally both indexes should be calculated in order to have a complete and informative US analysis.

We review details for both PRCC and eFAST in the next section. We tested the correctness of our Matlab implementation of LHS/PRCC and eFAST by running similar experiments with the softwares SaSat (see Hoare et al., 2008) and SimLab (2006), or comparing with known results for eFAST (see Ishigami function, pp. 41–42 in Saltelli et al. (2000) (data not shown).

3.1. *Partial rank correlation coefficient (PRCC)*

Correlation provides a measure of the strength of a linear association between an input and an output. A CC between x_j and y is calculated as follows:

$$r_{x_j y} = \frac{\mathrm{Cov}(x_j, y)}{\sqrt{\mathrm{Var}(x_j)\mathrm{Var}(y)}} = \frac{\sum_{i=1}^{N}(x_{ij} - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_{ij} - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}}, \quad j = 1, 2, \ldots, k \qquad (1)$$

and varies between -1 and +1. Cov(x_j, y) represents the covariance between x_j and y, while Var(x_j) and Var(y) are respectively the variance of x_j and the variance of y ($\bar{x}$ and $\bar{y}$ are the respective sample means). If x_j and y are the raw data, then the coefficient r is called sample or Pearson CC (Fig. 1C). If the data are rank-transformed, the result is a Spearman or rank correlation coefficient (Fig. 1C). It is important to note that the process of rank-transforming data assumes that sampled model inputs are real-valued or can adopt many possible values. If a parameter takes only integer values and the range of possible values it can assume is less than N, there is insufficient information to break ties during ranking, resulting in poor correlations. We are currently pursuing methods to handle this problem.[3]

Partial correlation characterizes the linear relationship between input x_j and output y after the linear effects on y of the remaining inputs are discounted. The PCC between x_j and y is the CC between the two residuals $(x_j - \hat{x}_j)$ and $(y - \hat{y})$, where $\hat{x}_j$ and $\hat{y}$ are the following linear regression models:

$$\hat{x}_j = c_0 + \sum_{\substack{p=1 \\ p \neq j}}^{k} c_p x_p \quad \text{and} \quad \hat{y} = b_0 + \sum_{\substack{p=1 \\ p \neq j}}^{k} b_p x_p \qquad (2)$$

Similarly to PCC, partial rank correlation (PRC) performs a partial correlation on rank-transformed data: x_j and y are first rank transformed, and then the linear regression models described in Eq. (2) are built. PRCC is a robust sensitivity measure for nonlinear but monotonic relationships between x_j and y, as long as little to no correlation exists between the inputs (see *Uncertainty and sensitivity functions and implementation* on our website http://malthus.micro.med.umich.edu/lab/usanalysis.html for the use of scatter plot functions to enable graphic detection of non-monotonicities).

By combining the uncertainty analyses with PRCC, we are able to reasonably assess the sensitivity of our outcome variable to parameter variation (see for example Blower and Dowlatabadi (1994) and Saltelli (2004). Fig. 2 shows an example of a standard LHS–PRCC scheme, scatter plots with correlation indexes (Pearson, Spearman, and PRCC, see Fig. 2C) and p-values (see the titles of the scatter plots in Fig. 2C) based on a classic ordinary differential equation (ODE) in population dynamics: a predator–prey (or Lotka–Volterra) model. The Lotka–Volterra model is the simplest model of predator–prey

[2] Definition of rank-transformation: the smallest value of a variable is assigned a rank of 1, the next largest value is assigned a rank of 2, tied values are assigned an average rank, and the largest value is assigned a rank equal to the sample size.

[3] The standard procedure when ties are encountered is to assign the tied values the average of what their values would have been if they have been consecutive but not equal. We are indirectly addressing the problem of integer-valued parameters with very few values by a single-parameter space exploration, holding them fixed during the UA and SA procedures.

interactions and was developed independently by Lotka (1925) and Volterra (1926):

$$\dot{Q} = \alpha Q(t) - \beta Q(t)P(t), \quad Q(0) = 10(\#\text{prey}) \tag{3}$$

$$\dot{P} = -\sigma P(t) + \delta Q(t)P(t), \quad P(0) = 5(\#\text{predator}) \tag{4}$$

It has two state variables (Q, P) and several parameters. Q represents the density of prey, P represents the density of predators, α is the intrinsic rate of prey population increase, β is the predation rate coefficient, σ is the predator mortality rate, and δ is the reproduction rate of predators per prey consumed. As an example, we assume that these four parameters follow normal pdfs (Fig. 2A) with means given by

$$(\alpha = 1.5, \quad \beta = 1, \quad \sigma = 3, \quad \delta = 1) \tag{5}$$

and initial conditions as shown in Eqs. (3) and (4). Standard deviations for parameters α and δ are set very small (i.e. std = 0.01) while parameters β and σ are varied in a larger range (i.e. std = 0.2). LHS is performed following the scheme

$$\begin{cases} \alpha \sim \text{Normal}(1.5, 0.01) \\ \beta \sim \text{Normal}(1, 0.2) \\ \sigma \sim \text{Normal}(3, 0.2) \\ \delta \sim \text{Normal}(1, 0.01) \end{cases} \tag{6}$$

The sample size N is set to 1000. Fig. 2A shows the cumulative distribution functions (CDFs) for each parameter, while Fig. 2B shows the outputs ($Q(t)$-prey and $P(t)$-predator) over time corresponding to the LHS scheme illustrated in Fig. 2A.

3.2. Inference on PRCCs

Significance tests can be performed to assess if a PRCC is significantly different from zero (thus, even small correlations may be significant) and if two PRCC values are significantly different from each other. Each PRCC (γ) generates a value T according to the following statistic (see Anderson, 2003, p. 143):

$$T = \gamma\sqrt{\frac{(N-2-p)}{1-\gamma^2}} \sim t_{N-2-p} \tag{7}$$

where T follows a student's t distribution with (N−2−p) degrees of freedom. N is the sample size and p is the number of inputs/parameters whose effects are discounted when the PRCC is calculated. For example, if we vary 6 inputs/parameters (x_i, i = 1,2,…,6) in LHS, p would be equal to 5 (PRCC(x_i, y) = γ_{x_iy/x_j}, $j = 1, \ldots, 6$ $j \neq i$). Eq. (7) is exact for linear partial correlation when the inputs and output are normally distributed, but is a large-sample approximation otherwise. Fisher showed that the following transformation of a sample Pearson CC r

$$r_1' = \frac{1}{2}\ln\left|\frac{1+r}{1-r}\right| \sim N\left(\mu, \frac{1}{\sqrt{N-3}}\right) \tag{8}$$

is normally distributed with mean equal to the unknown Pearson correlation of the population (μ) and standard deviation equal to $\left(1/\sqrt{N-3}\right)$. In order to test if two Pearson CCs are different, the log transformation in Eq. (8) is applied to the

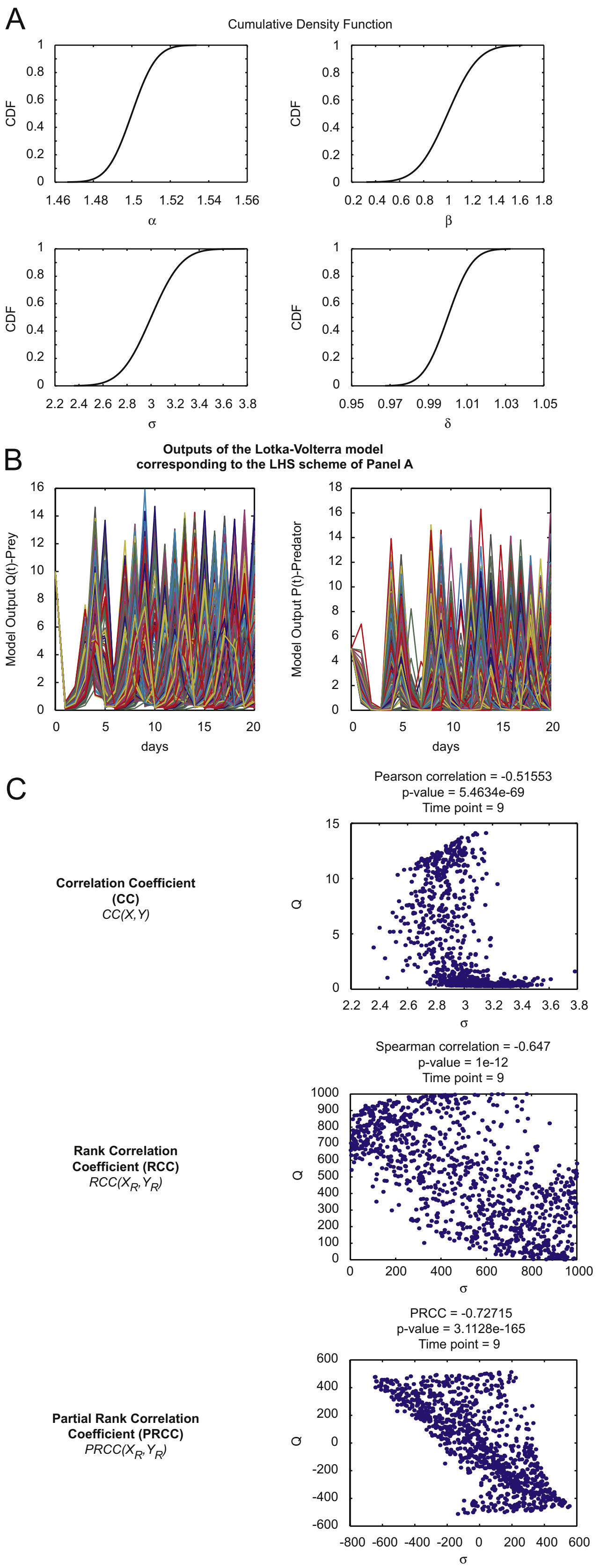

Fig. 2. General LHS scheme and PRCC performed on the Lotka–Volterra model (model equations and parameters are as described in Section 3.1). (A) Cumulative distribution functions—CDFs of 1000 samples independently drawn from normal pdfs (initialized following Eq. (6)) for the four parameters (α, β, σ, δ) of the Lotka–Volterra model described in Section 3.1, Eqs. (3) and (4). (B) Outputs of the Lotka–Volterra model over time (days) corresponding to the parameter combinations of the LHS scheme illustrated in (A). (C) Example of sampling-based correlation indexes calculated on the LHS matrix resulting from (A) and from the output matrix resulting from (B). The reference output is the variable $Q(t)$-prey at time t = 9 and it is shown on the ordinate. Parameter σ is taken as the parameter of interest and it is represented on the abscissa. Each dot represents the output value $Q(9)$ for a specific sampled value of parameter σ. Note that at each step of processing (correlation, rank correlation, partial rank correlation), the linear relationship between parameter and output variations becomes more apparent.

respective sample Pearson CCs r_1 and r_2 and the following statistic z is calculated (z-test):

$$z = \frac{r_1' - r_2'}{\sqrt{1/(N_1-3) + 1/(N_2-3)}} \sim N(0,1) \tag{9}$$

where N_1 and N_2 are the respective sample sizes. Since the distribution of sample PCCs is of the same form as that of Pearson CCs, the z statistic given in Eq. (9) can be used to compare them (after applying the log transformation (8) to the PCCs (pcc')):

$$z = \frac{pcc_1' - pcc_2'}{\sqrt{1/(N_1-3-p_1) + 1/(N_2-3-p_2)}} \sim N(0,1) \tag{10}$$

where N_1 and N_2 are the respective sample sizes and p_1 and p_2 are the respective inputs/parameters whose effects are discounted when pcc_1 and pcc_2 are calculated. The extension of Eq. (10) to PRCCs can be inferred from Anderson (2003, p. 144).

3.3. *Extended Fourier amplitude sensitivity test—eFAST*

eFAST, developed by Saltelli (2004), Saltelli and Bolado (1998), and Saltelli et al. (1999, 2000), is based on the original FAST developed by Collins and Avissar (1994), Cukier et al. (1973), and Schaibly and Shuler (1973). eFAST is a variance decomposition method (analogous to ANOVA): input parameters are varied, causing variation in model output. This variation is quantified using the statistical notion of variance $s^2 = \sum_{i=1}^{N}(y_i - \bar{y})^2 \big/ N-1$, where $N=$ sample size (or equivalently, total number of model runs), $y_i = i$th model output, and $\bar{y} =$ sample mean. The algorithm then partitions the output variance, determining what fraction of the variance can be explained by variation in each input parameter (i.e. partial variance). Partitioning of variance in eFAST works by varying different parameters at different frequencies, encoding the identity of parameters in the frequency of their variation (see Supplement A.1 online for details). Fourier analysis then measures the strength of each parameter's frequency in the model output. Thus, how strongly a parameter's frequency propagates from input, through the model, to the output serves as a measure of the model's sensitivity to the parameter.

The sampling procedure implemented in eFAST defines a sinusoidal function of a particular frequency for each input parameter (i.e. a *search curve*), $x = f(j)$, $j = 1,2,\ldots,N_S$, that assigns a value to x based on the sample number 1 through the total number of samples per search curve, N_S. The choice of sinusoidal function depends on the distribution of parameter values desired (e.g. uniform, normal, etc.). The frequencies assigned to parameters must meet several criteria so that they can be distinguished during Fourier analysis. See Supplement A.1 online for a detailed discussion of how search curves are specified and frequencies are chosen. The minimum recommended value for N_S is 65 (see Saltelli et al. 2000, p. 187). Due to the symmetry properties of trigonometric functions, the sinusoidal function will eventually repeat the same samples. A *resampling* scheme is implemented to avoid this inefficiency (Saltelli et al., 1999): eFAST algorithm is repeated N_R (the resampling size) times with different search curves specified by introducing a random phase shift into each sinusoidal function. So, the total number of model simulations, N, is given by $N = N_S \times k \times N_R$, where k is the number of parameters analyzed.

As an example, in Fig. 3, we illustrate the steps within the eFAST algorithm to the Lotka–Volterra model described in Section 3.1. We use 257 samples per search curve with no resampling (i.e. $N_S = 257$, $N_R = 1$). Parameter σ is taken as the parameter of interest. The algorithm assigns parameter σ a frequency of 31 and parameter β a frequency of 2. For simplicity, parameters α (frequency $=1$) and δ (frequency $=3$) are not illustrated. Fig. 3A illustrates this sampling step. Each sampled parameter combination is then used to solve the model (Fig. 3B).

The primary advantage of the eFAST method over the original FAST is the ability to calculate both the first-order sensitivity and total-order sensitivity of each input parameter (see Supplement A.2 online for details). A first-order sensitivity index, S_i, of a given parameter i, is calculated as the variance at a particular parameter's unique frequency (and harmonics of that frequency) divided by total variance (Fig. 3D, white pie-slice). First, variance (s_i^2) is calculated from the Fourier coefficients at the frequency of interest, j:

$$s_i^2 = 2(A_j^2 + B_j^2),$$

where

$$A_j = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(jx)\,dx,$$

$$B_j = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(jx)\,dx \tag{11}$$

then, the first-order S_i is calculated as a fraction of total variance:

$$S_i = s_i^2 / s_{total}^2 \tag{12}$$

This index represents the fraction of model output variance explained by the input variation of a given parameter. To estimate the total-order sensitivity index, S_{Ti}, of a given parameter i, eFAST first calculates the summed sensitivity index of the entire complementary set of parameters (i.e. all parameters except i) using their identification frequencies (Fig. 3D, black pie-slice).

S_{Ti} is then calculated as the remaining variance after the contribution of the complementary set, S_{ci}, is removed (Fig. 3D, gray pie-slice):

$$S_{Ti} = 1 - S_{ci} \tag{13}$$

This includes higher-order, nonlinear interactions between the parameter of interest and the complementary set of parameters. eFAST indexes can also be used to determine the degree of additivity of a model (see Supplement A.3 online for details).

3.3.1. *Inference on eFAST and the dummy parameter*

Eq. (7) allows statistical inference on PRCCs. Since an equivalent test for variance-based sensitivity indexes is not available, we propose a novel method based on dummy parameters for determining the significance of eFAST first- and total-order indexes. The use of dummy parameters is a standard practice in screening methods (see Saltelli et al., 2000, Chapter 4), although, to our knowledge, it has never been applied in the context of eFAST with the purpose of testing the significance of first- and total-order sensitivity coefficients.

eFAST implements random resampling (N_R) of search curves for more efficient parameter sampling. Because different search curves will produce different combinations of parameter values, different search curves will lead to slightly different sensitivity measures. We take advantage of these repeated measures to perform statistical tests comparing eFAST sensitivity indexes. The algorithm already performs these repeated measures, and thus the statistical tests entail no additional computationally intensive model simulations.

The eFAST method artifactually produces small but non-zero sensitivity indexes for parameters to which the model is completely independent. This is also true for the PRCC method. To illustrate this point, we run an example SA where we include a negative-control "dummy" parameter. This dummy parameter does not appear in the model equations and does not affect the model in any other way, so should ideally be assigned a sensitivity index of zero. Fig. 4 shows eFAST and PRCC SA results with the use

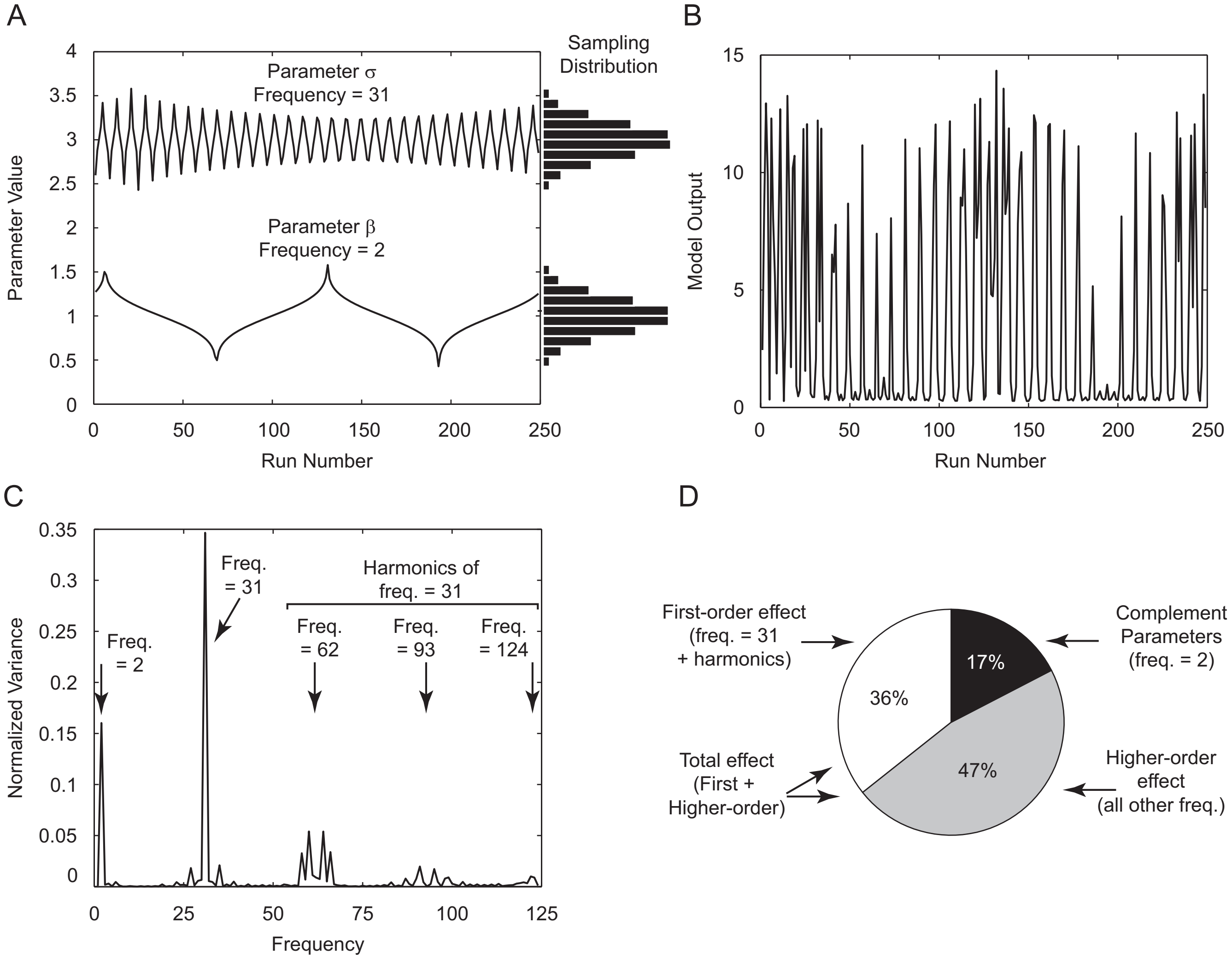

Fig. 3. eFAST performed on the Lotka–Volterra model. Model equations and parameters are as described in Section 3.1. (A) Input parameter sampling: parameters are varied according to a sinusoidal function based on run number. Note that the shapes of these search curves result in normal distribution of parameter values (see Eq. (6) for details) when sampled (horizontal histograms). σ is assigned a frequency of 31, and β is assigned a frequency of 2. α and δ are sampled at frequencies 1 and 3, respectively (not shown). These frequencies are chosen automatically to meet the criteria described in Appendix A.1. (B) Model output: the model is solved for each parameter combination from (A). The number of prey at $t=9$ is taken as the model output. Note that both high- and low-frequency components are evident by inspection. (C) Fourier analysis and variance spectrum: the variance at each frequency is calculated from the model output (see text) and normalized to total variance. The variance at frequency 2, 31, and higher harmonics of frequency 31 are indicated (arrows). (D) Sensitivity indexes: taking parameter σ (varied at frequency 31) as our parameter of interest, first-order sensitivity index (S_i) is calculated by the sum of variance at frequency 31 and higher harmonics, normalized to total variance (white pie slice). The sensitivity of the complementary set of parameters is calculated similarly (black pie slice). The remaining variance is assumed to be the result of non-linear higher-order interaction between parameters (gray pie slice). The total-order sensitivity index (S_{Ti}) is calculated by the sum of S_i and higher-order effects (white+gray pie slices).

of a dummy parameter on the Lotka–Volterra model discussed in Section 3.1.

The eFAST algorithm assigns this dummy parameter a small but non-zero first-order sensitivity index ($S_{dummy}=0.003$), and total-order sensitivity index (S_{Tdummy}) of approximately 0.11 (Fig. 4A). The non-zero first-order index, S_{dummy}, is likely derived from aliasing and interference effects (see Supplement A.1 online). The assignment of a larger artifactual value to the total-order index, S_{Tdummy}, is more complicated, as this artifact is derived from imprecise simplifying assumptions used to calculate the total-order index (see Supplement A.4 online). Since we cannot test if those values are significantly different from zero, by taking a dummy parameter as the parameter of interest, we propose a way to quantify these artifacts and test for significance. Parameters with a total-order sensitivity index less than or equal to that of the dummy parameter should be considered not significantly different from zero.

We propose using a *two-sample t-test*[4] on data generated by resampling the eFAST search curve to determine whether the sensitivity indexes of a parameter of interest are significantly different from the indexes calculated for the dummy parameter. The *t*-test compares two distributions, the S_i^j or $S_{T_i}^j$ ($j=1,2,\dots,N_R$) with the S_{dummy}^j or $S_{T_{dummy}}^j$ ($j=1,2,\dots,N_R$) (see Supplement A.5 online for more details).

Since N_R is usually small (for ease of computation), the adequacy of this *t*-test procedure is limited, considering that the normality assumption (see footnote 4) is unlikely to hold for the first- and total-order sensitivity indexes. However, it indicates

[4] We use the Matlab function *ttest2* to perform the *two-sample t-test*. We assume the most conservative options, e.g. two tails and that the two samples come from normal distributions with unknown and unequal variances. This is known as the Behrens–Fisher problem. *ttest2* uses Satterthwaite's approximation for the effective degrees of freedom.

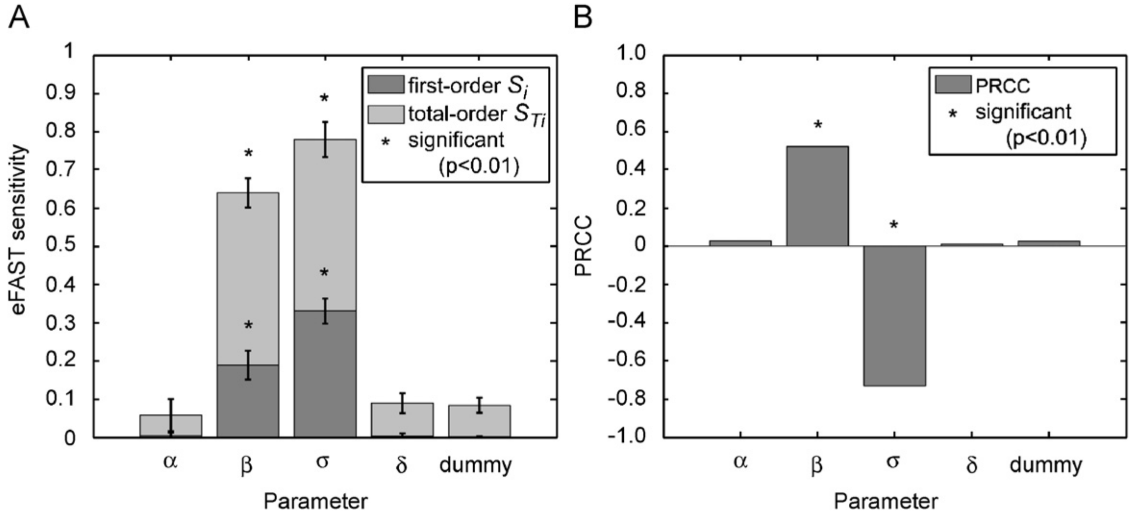

Fig. 4. Dummy parameter on eFAST and PRCC performed on the Lotka–Volterra model. Model equations and parameters are as described in Section 4.1 and Table 1. The reference output is $Q(t)$-prey Eq. (3), at $t = 9$. (A) eFAST results with resampling and significance testing. Search curves were resampled five times ($N_R = 5$), for a total of 1285 model evaluations ($N_S = 257$). First-order S_i and total-order S_{Ti} are shown for each parameter, including a dummy parameter, as described in the text. Error bars indicate ± 2 S.D. on the mean of resamples. Parameters with first- or total-order indexes significantly different ($p < 0.01$) from those of the dummy parameter are indicated with asterisks (*).(B) PRCC results. Sample size $N = 1000$. (*) denotes PRCCs that are significantly different from zero.

whether a small total-order sensitivity index should be considered an artifact of eFAST total-order estimation. The two-sample t-test might also be used to determine whether the indexes of multiple parameters of interest are significantly different from each other, allowing qualitative comparisons of effect size. Fig. 4A illustrates the use of resampling and statistical significance testing. The LHS/PRCC algorithm (Fig. 4B) assigns a PRCC value to the dummy parameter that is not significantly different from zero (0.04054, p-value >0.05 according to the statistic (7)).

Although not relevant for our examples, multiple test corrections should be considered if a large number of tests are performed in the US analysis (see Supplement B for details).

3.4. Sample size N

There is no *a priori* exact rule for determining the adequate sample size for either LHS–PRCC or eFAST. A minimum value is known for both LHS ($N = k$+1) and eFAST ($N_S = 65$), where k is the number of uncertain parameters that are varied. A way to overcome the problem is to systematically increase the sample size and check if the sensitivity index used (in our case either PRCC or eFAST S_i and S_{Ti}) can consistently capture and rank a similar set of most important effects. If that holds between two consecutive experiments, there is no evident advantage in increasing the sample size, because the conclusions (in terms of US analysis) will be the same. A measure of this type of correlation is given by the top-down coefficient of concordance (TDCC).

TDCC is described in Iman and Conover (1980, 1987) where a concordance measure is designed to be more sensitive to agreement on the top rankings from a set of h different rankings. It is based on Savage scores (Savage, 1956), defined as follows:

$$SS_i = \sum_{j=i}^{k} 1/j \tag{14}$$

where i is the rank assigned to the ith-order statistic in a sample of size k^5 (in our case k is the number of parameters varied).

If $h = 2$, the top-down correlation r_T is the simple CC computed on Savage scores. If $h>2$, r_T is the Kendall's coefficient of concordance (still computed on Savage scores). TDCC asymptotically follows a normal distribution, specifically $r_T\sqrt{n-1}{\sim}N(0,1)$. We implemented TDCC following Iman and Conover (1980, 1987), and it can be used for two purposes:

(i) selecting for an optimal sample size for LHS–PRCC and eFAST (see examples in Section 4)
(ii) comparing PRCC and eFAST to see if they are in agreement (not implemented).

An alternative method to assess the adequacy of sample size in LHS is based on the use of t-distribution with replicated sampling (see Sections 6 and 7 in Helton et al. (2000)).

The process of ranking in the PRC method refers to ranking the raw data from the LHS matrix and model output in descending order to calculate the PRC coefficient. Here and elsewhere, when we state that a set of parameters is the "most important" in affecting model output, we mean that after listing the sensitivity indexes in descending order by value for either PRCC or eFAST, these parameters are always greatest in absolute value. PRCCs can be formally ranked by the z-test (see Eq. (10)). We propose the use of a pairwise t-test and a dummy parameter to rank S_i and S_{Ti}.[6]

3.5. Time-varying sensitivity indexes

PRCC and eFAST indexes can be calculated for multiple time points and plotted versus time. This allows us to assess whether significances of one parameter occur over an entire time interval during model dynamics (this is the usual presentation of SA indexes for dynamical systems analysis). This analysis can be performed in two ways. If specific time points are known to be crucial, then only those will be checked for significance. For example, a model of acute virus dynamics will focus on the first days post infection. If we model a slow-growing pathogen (like *Mycobacterium tuberculosis*), we might be interested in later time points. If no particular time effect is known, the analysis can

[5] If $k = 3$, $SS_1 = 1 + \frac{1}{2} + \frac{1}{3}$, $SS_2 = \frac{1}{2} + \frac{1}{3}$, and $SS_1 = \frac{1}{3}$.

[6] The resampling size N_R should be large enough to achieve accuracy in statistical significance.

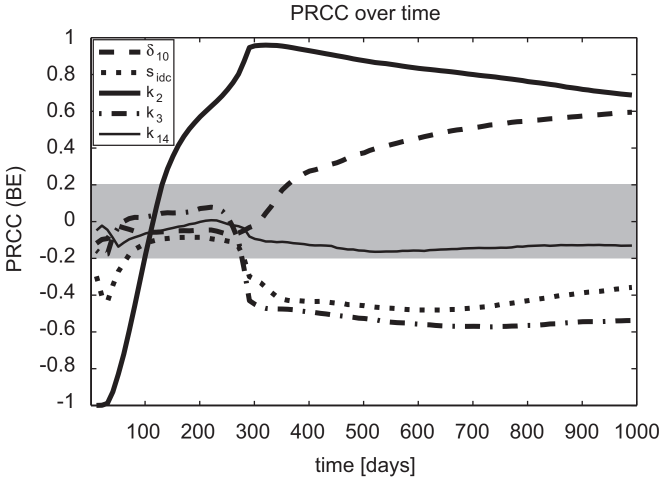

Fig. 5. PRCCs of the two-compartmental model US analysis performed in Section 4.3, plotted over a time course. PRCC is calculated with respect to extracellular bacteria load (BE), N = 200 (see Supplement E.2). The gray area represents PRCC values that are not statistically significant. Note how the sensitivity of parameters changes as system dynamics progress.

be done in an exploratory way, looking for any significant time-dependent relationships throughout the entire time course. Fig. 5 shows an example of PRCCs plotted over time: the US analysis is performed on the ODE model described in Section 4.3 (only five parameters are shown). The gray area in Fig. 5 indicates PRCCs that are not significantly different from zero (based on the statistic (7)). A similar plot can be displayed for eFAST indexes S_i and S_{Ti} (not shown).

In this example we see that the effect of parameter k_2 (maximum rate of infection of resting macrophages due to extracellular bacteria) changes with respect to bacterial load over time: it is negatively correlated (very strong PRCC, almost perfect negative correlation) right after infection (early time points), and then it becomes positively (very strongly) correlated as the infection progresses to its steady state. The positive sign of its PRCC indicates that if we increase parameter k_2, bacterial load increases (and vice versa). The negative sign suggests that if we increase it, bacterial load decreases (and vice versa). So, the rate of infection of resting macrophages by *M. tuberculosis* (k_2) is initially responsible for lowering the extracellular bacterial load (likely due to bacterial uptake). Then it becomes the most important source of infection, likely due to bacteria proliferation: the more bacteria internalized by macrophages, the more bacteria are released into the extracellular domain due to overproliferation and subsequent macrophage bursting or killing. Other mechanisms, such as the maximum rate of immature dendritic cells' (IDCs) activation/maturation/migration from the lung to the lymph node compartment (δ_{10}), only become significant later during infection (from 300 days on, see Fig. 5).

3.6. Standard versus explorative US analysis

US analysis can be implemented by choosing a specific reference output for calculating sensitivity indexes (for example, viral or bacterial load, if we track the progression of an infection). We define this type of approach a *standard* US analysis.

US analysis can be run on a list of several possible model outputs and the results can then be studied and classified depending on the goal of the analysis. We define this type of approach an *explorative* US analysis. For example, if the model comprises variables at different scales (e.g. intra-cellular versus extracellular, or molecular versus cellular) or in different compartments (e.g. different organs), we can choose outputs to investigate effects of parameter changes on different scales or in different compartments (see Section 4.3 for details). An intra-compartmental/intra-scale US analysis investigates how certain outputs generated in a specific compartment (scale) are affected by variations of parameters belonging to the same compartment (scale). An example of intra-compartmental US analysis can be to study how the rate of infection of resting macrophages in the lung affects the bacterial load in the lung (see Section 4.3).

Inter-compartmental US analysis explores how certain outputs generated in one compartment in a multi-compartment/scale model are affected by variations of parameters belonging to a different compartment/scale (for example, how the percentage of particular immune cells migrating from the lymph node—compartment 1—affects the bacterial load in the lung—compartment 2, see Section 4.3). The analysis can be performed either looking at specific inter- and intra-compartmental effects in which we are interested, or all the significant indexes can be listed and then classified as inter- or intra-compartmental effects. An example of multi-scale SA can be found in Chang et al. (2008) and Kirschner et al. (2007).

Whether a standard or an explorative US analysis is performed, the set of parameters that are varied is always under investigation and the parameters are continuously varied. The difference is in the goal of the analysis. Standard US analysis is applied for each of the examples shown in Section 4, while explorative US analysis is only applied to the two compartmental ODE model of *M. tuberculosis* infection (Section 4.3).

4. Uncertainty and sensitivity analysis examples

Since the relationship (including monotonicity) between parameters and outputs is not typically known *a priori*, then in principle using both PRCC and eFAST methods is ideal. The drawback is that issues related to accuracy of results and computational costs may arise. To illustrate the differences between these methods, we implement both PRCC and eFAST and compare the results for different types of mathematical models in biology: three different ODE systems (Lotka–Volterra, cell population dynamics in HIV infection, and *M. tuberculosis* infection), a delay differential equation (DDE) system in theoretical immunology, and an ABM of granuloma formation. The results are model-specific. However, in the discussion at the end of the manuscript, we suggest a general approach to balance accuracy of analysis and computational costs.

4.1. Emphasizing PRCC is not accurate when non-monotonicities are present: Lotka– Volterra model

We now revisit the predator–prey (or Lotka–Volterra) model described in Section 3.1. Our goal is to focus on two of the four parameters of model equations (3) and (4), and show how PRCC and eFAST give contradictory results, even with a simple prey–predator model. The CDFs of the parameter samples resulting from the LHS scheme are illustrated in Fig. 2A. We set the sample size N to 1000. Each parameter is independently sampled from normal pdfs (see Eq. (6)) and the model described by Eqs. (3) and (4) is simulated for each parameter combination. Fig. 6A shows the Lotka–Volterra model outputs (Eq. (6), $P(t)$) corresponding to the LHS matrix and scheme defined in Eq. (5). The vertical dashed line represents the time point chosen to perform the SA (time = 9 days). In order to test for nonlinearities and non-monotonicities between input variations (parameter σ) and output results ($P(t)$, predator), we produce scatter plots of the raw or ranked output

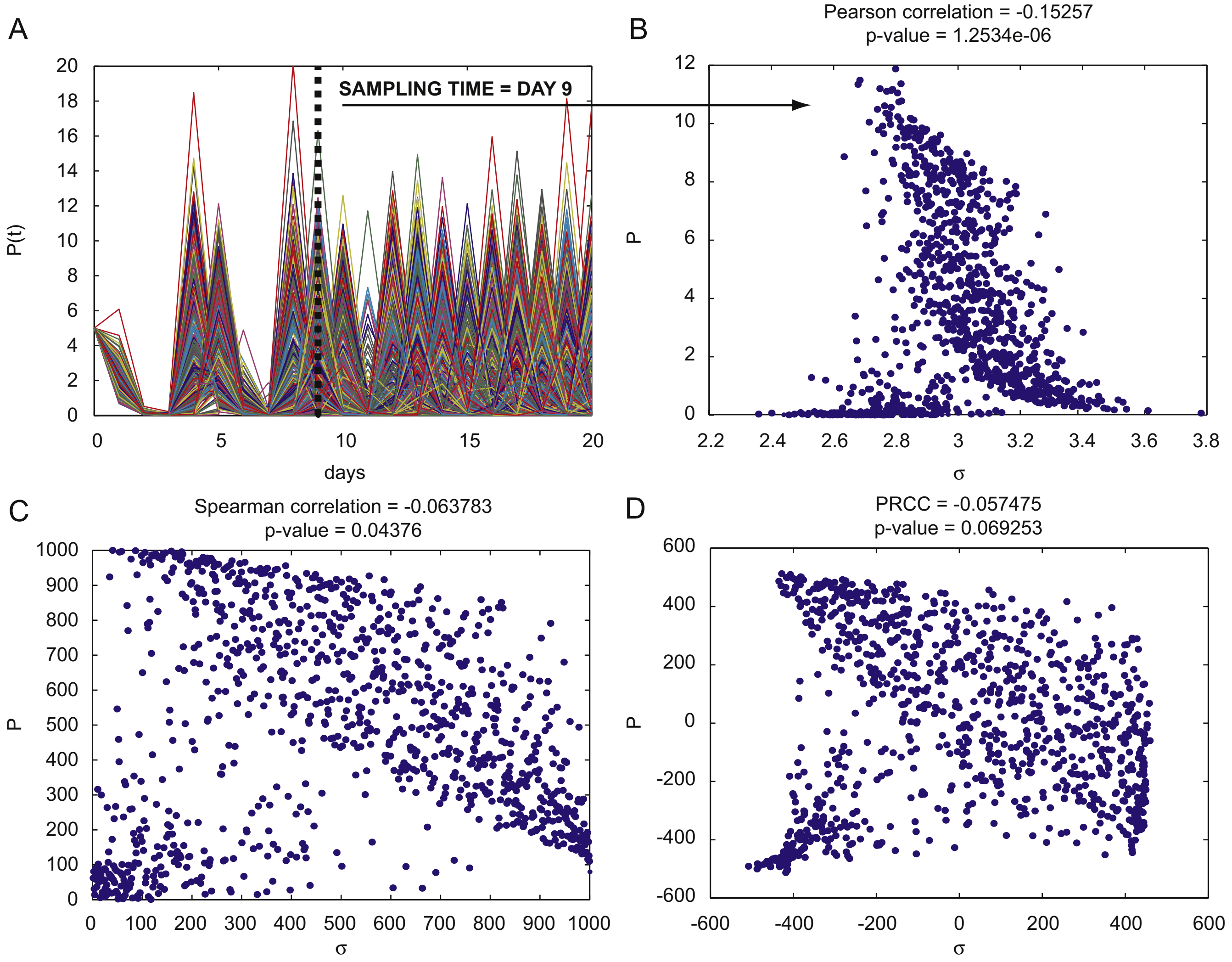

Fig. 6. (A) Plot of the output (1000 runs) of the Lotka–Volterra model. The ordinate represents variable $P(t)$, and the abscissa represents time (days). (B) Scatter plot (linear scale) of parameter σ (abscissa) and the "slice" of the output of (A) at time = 9 ($r_{Pearson} = -0.15257$, $p<0.001$). (C) Linear–linear plot of the rank-transformed data of (B) ($r_{Spearman} = -0.0638$, $p>0.04$, where $r_{Spearman}$ is the $r_{Pearson}$ coefficient calculated on the data of (C)). (D) Linear–linear plot of the residuals of the linear regressions of the parameter σ versus all the other parameters of the model (abscissa) and the residuals of the linear regression of the output versus parameter σ (ordinate) (PRCC = -0.0575, $p>0.06$). PRCC is the Pearson correlation coefficient calculated on the data of (D). Scatter plots of (B–D) remain qualitatively invariant for different LHS simulations. The input parameter (and its rank-transformed values) shown on the abscissa is σ, although all four parameters are varied simultaneously.

(ordinate) versus raw or ranked input (abscissa, see Panel B, C and D of Fig. 6).

Fig. 6B shows the linear scatter plot of the 1000 output values at day 9 plotted versus input variation. Fig. 6C shows the linear scatter plot of the rank-transformed data, while the scatter plot of two different residuals[7] used to calculate PRCC is shown in Fig. 6D (see figure legends for details). There is clearly a non-monotonic, nonlinear relationship between σ and the output $P(t)$ at time $t = 9$ (Fig. 6B–D) and the corresponding PRCC is not significantly different from zero (Fig. 6D, PRCC = -0.0575, $p>0.06$).

We designed the simulations varying parameters β and σ over large intervals, while parameters α and δ are allowed to vary around their respective mean value following a normal pdf with a very small standard deviation (e.g. 0.01). We tested for significant PRCCs for all four parameters of the Lotka–Volterra model. By construction, parameters α and δ should not significantly affect the output (as confirmed by PRCC columns in Table 1,

Table 1
Comparison of PRCC and eFAST values for Lotka–Volterra UA and SA (time point = day 9)

PRCC			eFAST			
			S_i (first order)		S_{Ti} (total order)	
Parameters	$Q(t)$	$P(t)$	$Q(t)$	$P(t)$	$Q(t)$	$P(t)$
α	0.0916*	0.0055	0.0042	0.0584	0.0584	0.0018
β	0.5586*	−0.1655*	0.1890*	0.6395*	0.6395*	0.0342*
σ	−0.7272*	−0.0575	0.3306*	0.7794*	0.7794*	0.2324*
δ	0.0422	0.0115	0.0032	0.0893	0.0893	0.0025
Dummy	0.0405	−0.0054	0.0013	0.0840	0.0840	0.0032

*Significant ($p<0.01$).
Values for the columns related to $Q(t)$ are illustrated in Fig. 4.

[7] We want to model y (response or dependent variable) as a function of x (regressor or independent variable), i.e. $y = f(x)$. The error we make in the prediction is called residual ($e = y - f(x)$). In our case x is the parameter vector (α, β, δ, σ) and f is a linear combination of a subset of x. Thus the x-axis of panel D represents [$e_x = \sigma - f(\alpha, \beta, \delta)$] and the y-axis represents [$e_y = y - f(\sigma)$], where the dependent and independent variables are rank-transformed.

Table 2
Parameter definitions and values of the HIV model described in Eqs. (15)–(18)

Parameter	Description	Range	Baseline
s	Rate of supply of CD4$^+$ T cells from precursors	[1e^{-2}, 50]	10 day^{-1} mm^{-3}
μ_T	Death rate of uninfected and latently infected CD4$^+$ T cells	[1e^{-4}, 0.2]	2e−2 day^{-1}
r	Rate of growth for the CD4$^+$ T-cell population	[1e^{-3}, 50]	0.03 day^{-1}
k_1	Rate constant for CD4$^+$ T cells becoming infected by free viruses	[1e^{-7}, 1e^{-3}]	2.4e−5 mm^3 day^{-1}
k_2	Rate latently infected cells convert to actively infected	[1e^{-5}, 1e^{-2}]	3e−3 day^{-1}
μ_b	Death rate of actively infected CD4$^+$ T cells	[1e^{-1}, 0.4]	0.24 day^{-1}
N_V	Number of free viruses produced by lysing a CD4$^+$ T cell	[1, 2e3]	1200
μ_V	Death rate of free virus	[1e^{-1}, 1e^1]	2.4 day^{-1}
T_{max}	Maximum CD4$^+$ T-cell population level	1500	1500 mm^{-3}

where all PRCCs for α and δ are between -0.05 and 0.05 and not significant, except for the PRCC of α for output Q (-0.0916, $p<0.004$).

Parameters β and σ should significantly affect the output: confirmed for the model output $Q(t)$ (first PRCC column of Table 1, 0.5586 and -0.7272, with p-values <0.001) but because of the non-monotonicity shown in Fig. 6D, PRCC of σ versus the output $P(t)$ is not significant (-0.0575, p-value >0.06).

On the other hand, the first-order (S_i) and total-order (S_{Ti}) sensitivity indexes returned by eFAST are the highest for σ (see Table 1, eFAST columns), suggesting that the contribution of σ to the variability of the output $P(t)$ is in fact the most important. According to condition (A.9) in Supplement A.5 online the sampling size for each curve ($N_S = 65$) ensures accurate S_i and S_{Ti} estimates (average $\mathrm{CV}_{\nabla S_i} < 6\%$ and $\mathrm{CV}_{\nabla S_{Ti}} < 2\%$ for both outputs).

Summary: These contradictory results show how, even with a simple model, nonlinear and non-monotonic relationships between input and output variation can lead to misleading conclusions during sensitivity analyses if PRCC is used.

4.2. Sampling (N_S) and resampling (N_R) in eFAST: an HIV–ODE model example

We next examine a model for the interaction of HIV with CD4+ T cells (see Perelson et al., 1993). It describes four cell population concentrations in the blood: uninfected T cells (T), latently infected T cells (T^*, i.e. cells that contain the provirus but are not producing new viruses), actively infected T cells (T^{**}, i.e. cells that are producing viruses), and free virus (V). The model comprises a total of nine parameters (see Table 2) and the dynamics of the various populations are given by

$$\frac{\mathrm{d}T}{\mathrm{d}t} = s - \mu_T T + rT\left(1 - \frac{T + T^* + T^{**}}{T_{max}}\right) - k_1 VT \tag{15}$$

$$\frac{\mathrm{d}T^*}{\mathrm{d}t} = k_1 VT - \mu_T T^* - k_2 T^* \tag{16}$$

$$\frac{\mathrm{d}T^{**}}{\mathrm{d}t} = k_2 T^* - \mu_b T^{**} \tag{17}$$

$$\frac{\mathrm{d}V}{\mathrm{d}t} = N_V \mu_b T^{**} - k_1 VT - \mu_V V \tag{18}$$

where $T(0) = 1000\,\mathrm{mm}^{-3}$, $T^*(0) = T^{**}(0) = 0$, and $V(0) = 10^{-3}\,\mathrm{mm}^{-3}$. The model admits two equilibrium solutions: an uninfected steady state (E_B, with no virus present) and an endemically infected steady state (E_P). N_V is a transcritical bifurcation parameter for E_B, or in other words, the system converges to the uninfected steady state E_B only if

$$N_V < N_{crit}, \quad N_{crit} = \frac{(k_2 + \mu_T)\mu_V + k_1 T_0}{k_1 k_2 T_0} \tag{19}$$

System (15)–(18) reaches the endemically infected steady state E_P if $N_V > N_{crit}$. The ranges over which parameters are varied (uniform pdfs) determine whether condition (19) is met or not. So, the parameters N_V, μ_T, μ_V, k_1, and k_2 (those included in (19)) should appear significant in our US analysis, depending on the intervals defined in Table 2 (*Range* column). Analytically, further conditions defined on other parameters play a role in determining the stability of E_P: whether they show up or not in the US analysis depends on the range over which they are varied (see Supplement F for an example on the effect of different parameter ranges on PRCC).

Parameter values given in Table 2 (*Baseline* column) produce a stable E_P, but stability can be lost through Hopf bifurcations occurring for several parameter combinations (see Perelson et al., 1993 for details).

We ran LHS/PRCC and eFAST analyses for different sample sizes. For LHS/PRCC we used the following sample sizes: 100, 200, 300, 400, 500, and 1000, while for eFAST we set N_S respectively to 65, 129, 257, 513, 1025, and 2049 (with $N_R = 5$). We test the adequacy of the sample size for both PRCC and eFAST with the TDCC. The number of parameters varied is 8, so the TDCC gives only a trend (since the statistic is only correct for large values of k, i.e. asymptotically). The results are summarized below, while all the details can be found in Supplement D online (Tables D.1–D.7).

4.2.1. PRCC and eFAST results for HIV model

PRCC results show how parameters N_V, k_2, and μ_v are consistently significant and the most important, with k_1 and μ_T always in fourth and fifth positions (see Table D.1 online). TDCC suggests $N = 200$ as the optimal sample size (see Table D.2 online).

As mentioned before, use of a uniform or log-uniform distribution (i.e. log-scale sampling) over a range spanning many orders of magnitude can produce very different results. In fact by using a log-scale sampling scheme, parameters k_1, N_V, and μ_T are now consistently significant and the most important, with μ_V always in fourth position (see Table D.3 online). Thus, by sampling on a log scale, PRCC results are very different. Only parameter N_V is still identified as significant, while parameters k_1 and μ_V lose significance. The TDCC suggests again $N = 200$ as the optimal sample size (see Table D.4 online).

eFAST results (both S_i and S_{Ti}) confirm that parameters μ_V, N_V, and k_2 are consistently significant and the most important (see Table D.5 online). The TDCCs suggest $N_S = 257$ as the optimal sample size, although increasing N_S does not improve the agreement in the results, except for $N_S = 2049$ (see Table D.7 online). The coefficients of variation (see Supplement A.5 online for implementation) are always very high for the first-order sensitivity indexes S_i (likely because the values of S_i are close to zero) and generally below 10–15% for the total-order sensitivity indexes S_{Ti} (see Table D.6 online for details). Increasing N_S only improves the coefficients of variation for S_{Ti}. We did not implement a log-scale sampling for eFAST.

Table 3
PRCC and eFAST results on the two-compartmental ODE model

Sensitivity index	Parameters with significant sensitivity index		
(A) *US analysis with extracellular bacterial load (BE) as output of interest over three time points*			
	Day 100	Day 500	Day 1000
PRCC	k_2, $s_{IDC}(-)$, $k_{14}(-)$	k_2, $k_3(-)$, $s_{IDC}(-)$, δ_{10}, k_4, $k_{14}(-)$	k_2, δ_{10}, $k_3(-)$, $s_{IDC}(-)$, k_4, $k_{14}(-)$, $\xi(-)$, δ_8
eFAST$-S_i^{(*)}$	$\boldsymbol{k_2}$	$\boldsymbol{k_{14}}$, ξ, k_2, k_3, s_{IDC}	$\boldsymbol{k_3}$, $\boldsymbol{k_{14}}$, k_2, s_{IDC}, ξ
eFAST$-S_{Ti}^{(*)}$	$\boldsymbol{k_2}$	$\boldsymbol{k_{14}}$, ξ, k_2, k_3, s_{IDC}, k_4	$\boldsymbol{k_3}$, $\boldsymbol{k_{14}}$, k_2, s_{IDC}, ξ, k_4
(B) *US analysis with extracellular bacterial load (BE) as output of interest, at time 1000 days post infection*			
	Inter-compartment	Intra-compartment	
PRCC	δ_{10}, $\xi(-)$	k_2, $k_3(-)$, $s_{IDC}(-)$, k_4, $k_{14}(-)$, δ_8	
eFAST$-S_i^{(*)}$	ξ	$\boldsymbol{k_3}$, $\boldsymbol{k_{14}}$, k_2, s_{IDC}	
eFAST$-S_{Ti}^{(*)}$	ξ	$\boldsymbol{k_3}$, $\boldsymbol{k_{14}}$, k_2, s_{IDC}, k_4	
(C) *US analysis with mature dendritic cells (MDCs) as parameter of interest, at time 500 days post infection*			
	Inter-compartment	Intra-compartment	
PRCC	$\boldsymbol{s_{IDC}}$, $\boldsymbol{k_2}$, $k_3(-)$, $\delta_9(-)$	$\boldsymbol{\delta_{10}}$	
eFAST$-S_i^{(*)}$	$\boldsymbol{k_{14}}$, k_4, k_3, s_{IDC}	$\boldsymbol{\xi}$	
eFAST$-S_{Ti}^{(*)}$	$\boldsymbol{k_{14}}$, k_4, k_3, s_{IDC}, k_2	$\boldsymbol{\xi}$	
(D) *US analysis with T helper cells type I (Th1) as output of interest, at time 100 days post infection*			
	Inter-compartment	Intra-compartment	
PRCC	$\boldsymbol{\xi}$, $\boldsymbol{\delta_{10}}$	$\boldsymbol{k_3}$, k_2, s_{IDC}, δ_4, $k_{14}(-)$	
eFAST$-S_i^{(*)}$	$\boldsymbol{\delta_{10}}$	$\boldsymbol{k_2}$, $\boldsymbol{s_{IDC}}$	
eFAST$-S_{Ti}^{(*)}$	$\boldsymbol{\delta_{10}}$	$\boldsymbol{k_2}$, $\boldsymbol{s_{IDC}}$, δ_8	

This table summarizes results shown in great detail in Supplement E.2 and E.3 online where different sample sizes are used for both LHS/PRCC and eFAST. The time points tested are [100, 500, 1000]. The sign of PRCC is shown in parenthesis. The top two or three parameters are in bold italics. The order they are listed reflect the magnitude of the coefficient (in absolute value, going from high to low). (*): Significant; i.e. $p<0.01$. (A) PRCC and eFAST results grouped by significance over time. Extracellular bacterial load (BE) is the output of interest. *Panel* (B) Inter- and intra-compartmental effects. PRCC and eFAST results with BE as output of interest, at time 1000. (C) Inter- and intra-compartmental effects. PRCC and eFAST results with MDC as output of interest, at time 500. (D) Inter- and intra-compartmental effects. PRCC and eFAST results with Th1 as output of interest, at time 100.

If we combine results from both PRCCs and eFAST S_i and S_{Ti}, parameters μ_T, k_1, k_2, N_V, and μ_V indeed are significant, suggesting how a properly designed US analysis should return bifurcation parameters as significant. Note that in the original paper (Perelson et al., 1993) these were shown analytically to be bifurcation parameters.

Summary: PRCC and eFAST measure different things. This example shows how results can overlap for a subset of parameters (e.g. μ_V, N_V, and k_2). However, the ranking of the most important effects is not preserved between the two methods. The TDCC shows how PRCC more efficiently (i.e. with a lower sample size) achieves a concordance in ranking the most important effects, while eFAST may need larger sample sizes.

4.3. Standard and explorative US analysis: a two compartmental ODE model

We now examine a more complex ODE to address the use of US analysis in a multi-scale/multi-compartmental setting. As an example, we choose an ODE model of *M. tuberculosis* infection in humans published by our group (Marino and Kirschner, 2004): it comprises a total of 17 equations, and a much larger set of 90 parameters that describe dynamics between two physiological compartments (lymph node and lung). A detailed description of the model can be found in Marino and Kirschner (2004) and Marino et al. (2004). Basic details on the biology are given in Supplement E and Fig. E.1.

We perform US analysis on the 12 parameters that are specifically involved in establishing infection and linking the two anatomical compartments (see Table E.1 online for details). This set of parameters represents mechanisms whose action is elicited either in the lung or in the lymph node compartment. A more comprehensive analysis would include all the parameters of the model, or at least the uncertain ones. Here we focus on a subset for the purpose of illustrating both a standard and an explorative US analysis. As discussed in Section 3.6, whether a standard or an explorative US analysis is performed, the set of 12 parameters varied is always under investigation and the parameters are continuously varied. The difference between the two analyses is subtle. A standard US analysis lists all the parameters with a significant sensitivity index with respect to one or more outputs of reference (in our example the output of reference is the extracellular bacterial load—BE, see Table 3, panel A).

An explorative US analysis identifies which one of these effects (significant sensitivities) has an impact in the same compartment (intra) of the output of reference, or in a different compartment (inter). In our example, we examined two outputs in the lung compartment (extracellular bacterial load—BE, type I T helper cells—Th1) and one output in the lymph node compartment (mature dendritic cells—MDCs) (see Supplement E for details on the biology). Then we computed both PRCC and eFAST coefficients and classified all the parameters with significant sensitivity indexes either as intra- or inter-compartmental effects, as having effects, depending on the output of reference.

We performed LHS/PRCC and eFAST analyses for different sample sizes. For LHS/PRCC we used the following sample sizes 100, 200, 300, 400, 500, and 1000, while for eFAST we set N_S respectively to 65, 129, 257, 513, 1025, and 2049 (with $N_R=5$). We tested the adequacy of the sample size for both PRCC and eFAST with the TDCC. The number of parameters varied is small (i.e. 12), so TDCC gives only a trend (as in the previous example). Since analytical results are not available due to the size of the model and the complexity of the equations, conclusions are not as straightforward as in the previous examples. Standard US analysis results are shown in Table 3, panel A, while an example of inter- and intra-compartmental analysis at specific time points is given in Table 3, panels B–D. All the details can be found in Supplement E.2 and E.3 online. We analyze three different time points, namely 100, 500, and 1000 days post infection. Table 3 shows both PRCC and eFAST S_i and S_{Ti}; they are listed in descending order (the first listed have the largest absolute values and are in bold italics) and the sign of the PRCCs is in parenthesis.

Overall, this example shows how the sets of significant S_i and S_{Ti} returned by efAST are consistent throughout the analysis, returning the same list of important parameters with the same ranking. eFAST results suggest sets of significant parameters that are not always listed as significant by PRCC and are generally smaller. Moreover, when the same set of parameters is returned, the ranking is frequently different between the two methods (see panel A).

Table 3 shows how the fraction of precursor T helper cells (Th0) migrating out of the lymph node compartment into the blood (ξ) and the maximum rate of IDCs' activation/maturation/migration from the lung compartment to the lymph node compartment (δ_{10}) are alternatively classified as significant inter- (panels B and D) or intra-compartmental (panel C) parameters by PRCC and eFAST.

However, the parameters are not consistently found significant by both indexes. For example, in panel C, δ_{10} is found significant only by PRCC, whereas parameter ξ is found as significant only by eFast: Each parameter is ranked as the most important by one method, but not identified at all by the other method. Another example is given by parameter ξ in panel D: only PRCC lists it as significant (while δ_{10} is returned as very important by PRCC and eFAST S_i and S_{Ti}).

Summary: This example shows how on a large complex model, the use of both indexes is recommended because often the set of the most important parameters is not consistent between the two methods.

4.4. A delay differential model example

We next apply US analysis to a DDE model published by our group (see Marino et al., 2007 for details). The model investigates the role of delays in innate and adaptive immunity to intra-cellular bacteria infection. It tracks five variables: uninfected target cells (X_U), infected cells (X_I), bacteria (B), and phenomenological variables capturing innate (I_R) and adaptive (A_R) immunity. A detailed description of the model is given in Marino et al. (2007). Here we only show the equations:

$$\frac{dX_U}{dt} = s_U - \alpha_1 X_U B - \mu_{X_U} X_U \tag{20}$$

$$\frac{dX_I}{dt} = \alpha_1 X_U B - \alpha_2 X_I A_R - \mu_{X_I} X_I \tag{21}$$

$$\frac{dB}{dt} = \alpha_{20} B\left(1 - \frac{B}{\sigma}\right) - \alpha_3 B I_R - \alpha_4 B A_R \tag{22}$$

$$\frac{dI_R}{dt} = s_{I_R} + \int_{t-\tau_1}^{t} w_1(s) f_1(B(s), I_R(s))\,ds - \mu_{I_R} I_R \tag{23}$$

$$\frac{dA_R}{dt} = s_{A_R} + \int_{t-\tau_2}^{t} w_2(s) f_2(B(s), A_R(s))\,ds - \mu_{A_R} A_R \tag{24}$$

Two delays are included in the model. The delay for innate immunity, τ_1, occurs on the order of minutes to hours and τ_2 is the delay for adaptive immunity on the order of days to weeks. We assume that both responses are dependent solely on the bacterial load ($f_1 \equiv f_2 = B(s)$) in the previous τ_i time units ($i = 1,2$), where the kernel functions $w_i(s)$ ($i = 1,2$) weight the past values of the bacterial load, $B(s)$.

We use a uniform kernel for innate immunity (with $\tau_1 \cong 1$) and an exponential growth kernel for adaptive immunity (with $\tau_2 \cong 20$). System (20)–(24) is comprised of 20 parameters (15 independent parameters), seven of which are directly involved in the existence condition for equilibrium solutions. Initial conditions for the model are given in Table G.1 (see Supplement G online). We solved the system numerically using the Matlab solver *dde23*.[8] The model admits two equilibria: boundary equilibrium E_B (clearance of infection, no bacteria) and a positive equilibrium E_P (damped or sustained oscillatory bacterial levels, depending on the values of τ_1 and τ_2). The positive equilibrium E_P exists only if the following condition holds:

$$\alpha_{20} - \alpha_3 \frac{s_{I_R}}{\mu_{I_R}} - \alpha_4 \frac{s_{AI_R}}{\mu_{A_R}} > 0 \tag{25}$$

We sampled eight parameters simultaneously (τ_1, τ_2, α_1, α_2, α_3, α_4, α_{20}, σ), defining uniform pdfs for their distributions. The remainder of the parameters are held constant at their default values (see Table G.1 online). Table G.1 is used to initialize the sampling procedure (see Value column) and to define intervals for UA (see Range column). We set the dimension of the sample in LHS to $N = 1000$, while eFAST is performed for three different N_S (i.e. 65, 129, and 257) and resampling size to $N_R = 5$ (for inference on S_i and S_{Ti}), except for the sample $N_S = 65$ where we increased N_R to 20 to improve accuracy.

Some parameter combinations do not satisfy the existence condition for E_P (i.e. condition (25)); thus, we calculate PRCCs both on the entire LHS dataset and on the subset of samples fulfilling condition (25). In contrast, the same procedure cannot be applied to eFAST. Because of its unique sampling technique, eFAST must process the whole sampled dataset at once, and we can only compare the sensitivity indexes calculated on the whole parameter space.

We summarize the main US analysis results below, while all the details can be found in Tables G.2 and G.3 online. The indexes are evaluated at five different time points (10, 30, 50, 100, and 200 days) and the model output chosen for SA is bacterial load (Eq. (22)). Condition (A.9) is never satisfied by the choice of N_S and N_R (see Table G.3 online). Panels A–C in Table G.2 online show PRCC results, while panels D and E show eFAST results. Both PRCC and eFAST suggest that bacterial load levels over time are mainly affected by variations in the parameters α_3 and α_{20}. PRCC scatter plots of these two parameters versus the bacterial load are shown in Fig. 7 (using the whole LHS matrix), where the strong correlations are confirmed. Parameter α_4 is found consistently significant only with the PRCC analysis (although the correlation is weak).

eFAST (Table G.2 online Panels D and E, online) and PRCC (Table G.2 online, Panels A and C) list delays τ_1 and τ_2 as important sources of variation for bacterial load (PRCC predicts they are significant with small negative correlations). Both delays are significant only with the subset of the LHS matrix satisfying (25) (Table G.2 online, Panel C online), in line with analytical results where τ_1 and τ_2 play a role only when E_P exists (τ_1 and τ_2 define the nature of the equilibrium, with either sustained or damped oscillations). The effect of τ_2 is lost if the whole LHS matrix is used (Table G.2, panel A, online) and both are lost (Table G.2 online Panel B) if only the subset of the LHS matrix not satisfying (25) is used (in fact E_B is not affected by changes in τ_1 and τ_2). PRCC scatter plots (see Fig. 7) do not show any clear nonlinear nonmonotonic relationship between variations in the parameters τ_1,

[8] Version 6.5, R13, Copyright 1984–2002, The MathWorks, Inc.

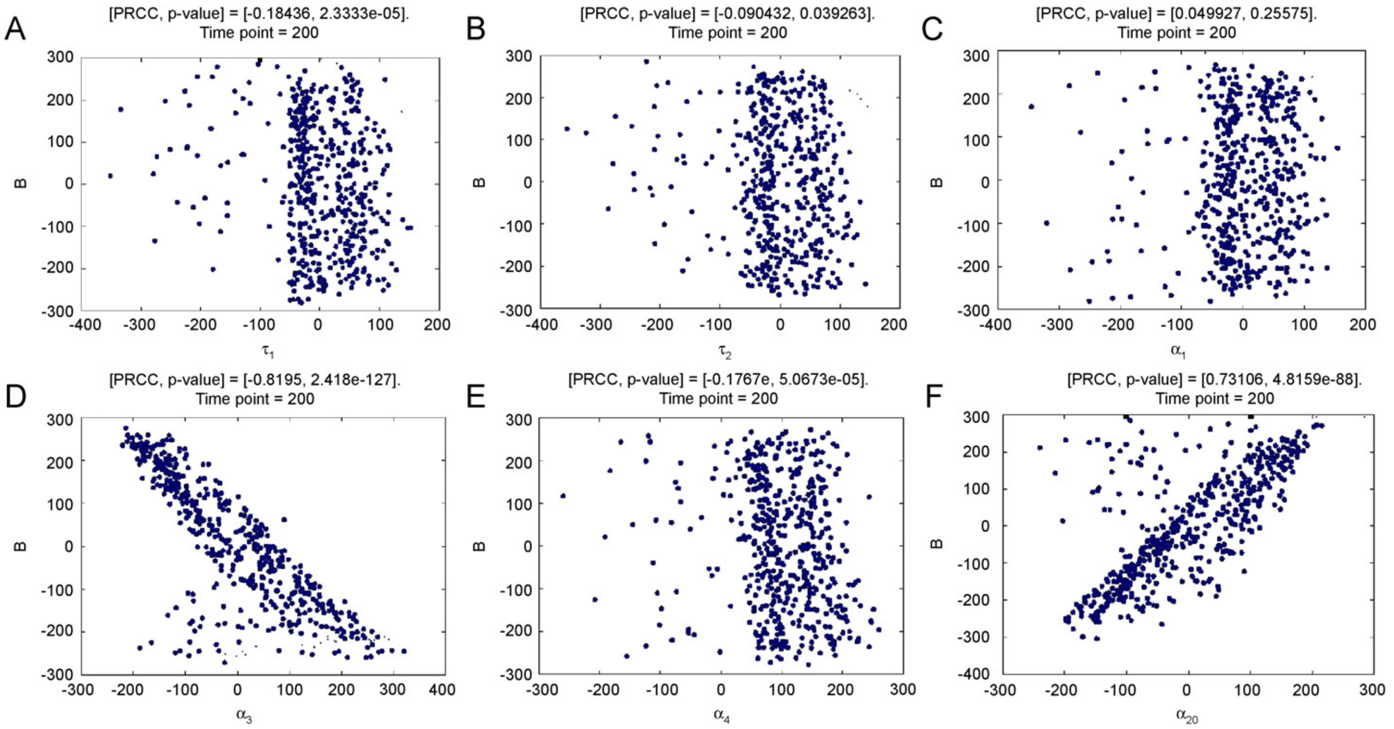

Fig. 7. PRCC scatter plots of parameters τ_1, τ_2, α_1, α_3, α_4, and α_{20} (calculated at day 200, all eight parameters are varied simultaneously). Sample size $N = 1000$. The abscissa represents the residuals of the linear regression between the rank-transformed values of the parameter under investigation versus the rank-transformed values of all the other parameters. The ordinate represents the residuals of the linear regression between the rank-transformed values of the output versus the rank-transformed values of all the parameter under investigation The title of each plot represents the PRCC value with the corresponding *p*-value (see Table G.2 at 200 days in Supplementary material online).

τ_2, α_1, and the output (bacterial load). PRCC and eFAST confirm indirectly our analytical results: they return as significant those parameters that are involved in the existence condition (25) for equilibrium solutions.

Summary: This example shows how PRCC and eFAST results can overlap. However, since the nature of the relationship between inputs and output is not known *a priori*, PRCC and eFAST should be used together (as in this example), if eFAST computational cost is not prohibitive.

5. Uncertainty and sensitivity analysis in agent-based models

ABMs (also called ''individual-based models'') are a formalism evolved from early research in cellular automata and artificial life. The defining feature of ABMs is that elements of the system are represented as discrete agents that move and interact according to defined rules, in an explicitly defined spatial environment. Stochasticity enters the model as some decision-making rules can be based on random chance, such as a random walk movement of cells. In an ABM, the individual, possibly stochastic, interactions between agents give rise to global, system-wide dynamics and patterns. Thus, ABMs are ideal for studying complex systems in which stochasticity, and spatial and temporal heterogeneity are important, such as biological systems.

Stochastic models, such as the ABM we consider here, pose unique challenges to US analysis. Most US analysis techniques have been developed for use with deterministic models, such as those presented in previous sections of this work. To our knowledge, very few researchers have attempted extensive and systematic US analysis on ABMs (Lempert et al., 2002; Riggs et al., 2008; Segovia-Juarez et al., 2004).

To perform US analysis on an ABM, it is critical to keep in mind the distinction between aleatory and epistemic uncertainty during model building. In this analysis, epistemic uncertainty is handled by holding a parameter to a fixed value during a particular model simulation, but allowing probabilistic variation between model simulations to reflect uncertainty. In contrast, stochastic components vary randomly from moment to moment within a single ABM simulation. Some studies have applied a stochastic component to otherwise deterministic models by sampling a random sequence of values prior to running a model simulation (Helton, 1999; Helton and Breeding, 1993; Helton et al., 1995). However, this approach cannot be easily applied to ABMs as presented here: stochastic decisions made by each agent at each time step are based on conditional probabilities that depend on both random chance and the state of other interacting agents. Therefore it is impossible to correctly specify a random sequence of agent decisions prior to running the simulation. When performing UA on an ABM, epistemic uncertainty and aleatory uncertainty become conflated in the model output. It is difficult to know whether variability in model outcome is due to experimentally introduced variation in input parameters (epistemic uncertainty) or to the inherent stochastic components of the model (aleatory uncertainty). Therefore, it is difficult to apportion variability to input parameters during SA. In this section, we perform US analysis on an ABM, identify the strengths and weaknesses of the PRCC and eFAST techniques in dealing with aleatory uncertainty, and propose an averaging method to reduce the influence of aleatory uncertainty.

5.1. An agent-based model example

As an example, we present a published ABM describing granuloma formation during *M. tuberculosis* infection (Segovia-Juarez et al., 2004). This was the first ABM to use a modification of a US analysis using PRCC. In this model, the spatial

Table 4
Parameters analyzed in US analysis

Symbol	Description	Range of uniform pdf
λ	Chemokine diffusion coefficient	[0.4, 0.8]
δ	Chemokine degradation coefficient	[2.8811×10^{-4}, 0.0011]
α_{BI}	Intra-cellular growth rate	[0.002, 0.006]
T_{recr}	Probability of T-cell recruitment	[0.10, 0.40]
T_{move}	Probability of T-cell movement	[0.01, 0.20]
T_{actm}	Probability of T-cell activating macrophage	[0.05, 0.20]
M_{init}	Initial number of macrophages	[40, 400]
M_{recr}	Probability of macrophage recruitment	[0.20, 0.70]
M_{asp}	Activated macrophage speed	[200, 8000]
N_{tact}	Number of T cells needed to activate macrophage	[1, 6]
p_k	Probability of bacteria killed by resting macrophage	[0.01, 0.10]
p_{Tk}	Probability of T cell kills a macrophage	[0.01, 0.10]

environment represents a 2 mm^2 section of lung tissue. This area is subdivided into a 100×100 lattice of 20 μm^2 micro-compartments. Agents (cells) populate this environment, representing resident or infiltrating macrophages and T cells. Bacteria and chemokines also exist in the model, but are treated as a continuous quantity rather than discrete agents. Each micro-compartment can contain at most one macrophage and one T-cell agent (and a quantity of bacteria/chemokines, as these are assumed to have negligible size). Some micro-compartments are designated as vascular source compartments, through which new macrophage and T-cell agents arrive from the blood.

A simulation is initiated by distributing an initial population of resting macrophages randomly on the lattice and an initial load of extracellular bacteria in the center of the lattice. The simulation evolves by a series of discrete time steps, during which rules are evaluated governing the diffusion of chemokines, infection and replication of bacteria, and arrival, movement, interaction, and change in phenotype of immune cells. Parameters that we will analyze by US methods control initial conditions and rates or probabilities used to evaluate rules (Table 4). See Segovia-Juarez et al. (2004) for a more detailed description of the model.

5.2. Aleatory versus epistemic uncertainty in ABMs

It is impossible to entirely separate model output variability due to either aleatory or epistemic uncertainty. However, it is useful to crudely quantify aleatory uncertainty in the absence of variability due to epistemic uncertainty by repeatedly solving the model, holding all parameters constant. Fig. 8 represents two typical scenarios: containment of infection and dissemination. The baseline set of parameter values differs between the two scenarios (see Segovia-Juarez et al., 2004) for definitions of these scenarios and baseline parameters). The variability around each scenario is determined by aleatory uncertainty, while epistemic uncertainty causes the emergence of this bimodal outcome. We find that the containment scenario produces a moderate level of aleatory uncertainty. To roughly quantify aleatory uncertainty, we calculate the coefficient of variation of 10 replicates, holding all parameters constant at their average values (middle value of the Range column in Table 4). For this model, coefficients of variation of the output for the containment scenario (extracellular bacterial load) are 3.5% at day 500 and 62.4% at day 30. The coefficient of variation at day 30 is much larger than at day 500, meaning that aleatory uncertainty at this early time point will more strongly mask any variability due to epistemic uncertainty introduced in subsequent analysis steps. We will focus our analysis on day 30, as this time point is analyzed in Segovia-Juarez et al. (2004), providing a reference point to compare our results, and because aleatory effects should be most problematic at this time point due to the high coefficient of variation.

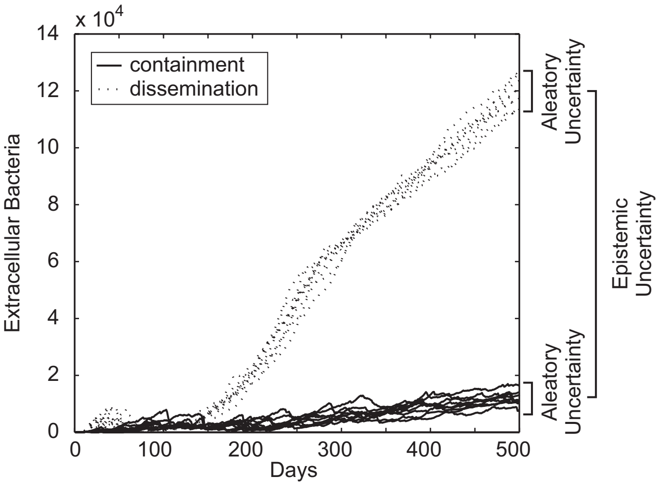

Fig. 8. Variation in ABM output due to epistemic uncertainty and aleatory uncertainty. Differences in parameter values lead to either containment of extracellular bacterial load (solid lines), or dissemination of bacteria (dotted lines), as described previously in Segovia-Juarez et al. (2004). Within each parameter combination, inherent stochasticity within the model causes aleatory uncertainty. At time points earlier than approximately day 150, aleatory uncertainty masks the epistemic difference that USA methodologies seek to measure.

5.3. LHS/PRCC results: averaging replicates can reduce computational cost

To study aleatory uncertainty, we propose a replication and averaging scheme that can be applied to either PRCC or eFAST analysis. First, parameter combinations are created using the LHS or eFAST internal sampling algorithm. Next, multiple model simulations are repeatedly run for each parameter combination (replication step). Note that when using a pseudo-random number generator in the algorithm, we reinitialize the random seed for each model simulation. Finally, the sensitivity coefficients are calculated using the average of model outputs across replicates. By comparing these results with the sensitivity coefficients obtained from each replicate individually, we can assess whether US analysis is robust at a specific sample size. One caveat to this approach is that using the average to characterize the distribution of output values over aleatory uncertainty is reliable only if the output values are clustered around a central value (i.e. unimodal). We apply the replication and averaging scheme to the Segovia-Juarez ABM.

5.4. LHS/PRCC results for ABM

When aleatory uncertainty is moderate, as is in this example, LHS/PRCC produces consistent results at large LHS sample size: in the original analysis of this model, Segovia-Juarez et al. use an LHS sample size of 1000. As sample size is reduced to less than approximately 200–300, two effects occur: first, PRCC becomes less reliable due to the influence of aleatory uncertainty; second, statistical power for PRCC significance is reduced due to the smaller sample size; therefore weakly correlated parameters are not found to be significant.

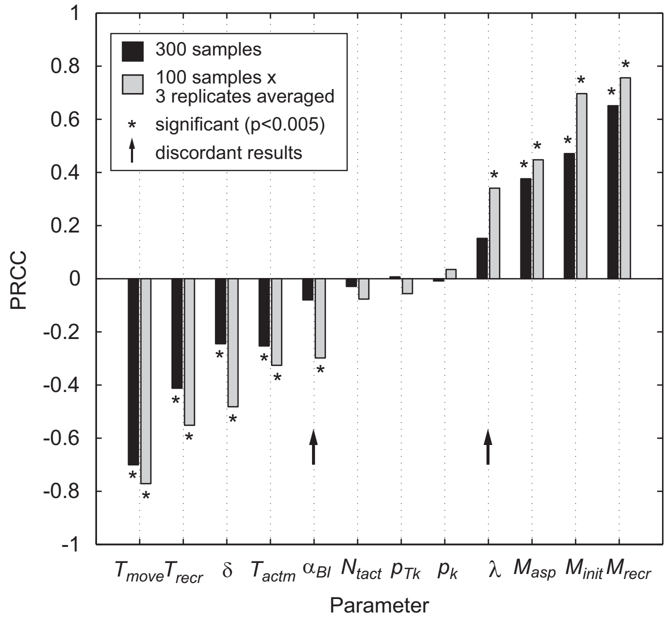

Fig. 9. Effect of aleatory uncertainty on LHS/PRCC and the benefit of averaging replicates. Model parameters were sampled 300 (black bars) or 100 times (gray bars). For the 100-sample condition, the three replicate model simulations were performed for each sample and averaged. Parameters with a PRCC significantly ($p<0.005$) different from zero are indicated with (*). Parameters with discordant statistical significance between the two experimental conditions are indicated (arrow).

Next, we explore whether averaging replicates can reduce effects of aleatory uncertainty, allowing fewer computationally intensive model simulations. We find that averaging replicates has two beneficial effects. First, PRCC becomes reliable at smaller sample sizes due to the removal of some of the variability induced by aleatory uncertainty. Specifically, by using replicates (three replicates in this case), a similar set of parameters with significant PRCC is achieved with a smaller number of samples (at least 50) compared to the standard scheme without replicates (at least 300 samples). Second, parameter values and model output becomes more tightly correlated due to the removal of some confounding aleatory noise, resulting in PRCC of greater magnitude (see Fig. 9). This increase in PRCC magnitude that occurs by averaging is apparent even when PRCC is already reliable due to large sample size, e.g. 1000 samples (not shown). These benefits come at a cost, however: if the total number of model simulations is held constant, performing replicates reduces the number of samples available for statistical tests (e.g. 300 parameter samples simulated once, versus 100 parameter samples replicated three times, then averaged, yielding 100 sample averages). In Fig. 9, we show that there are some situations where the gain in PRCC magnitude due to averaging is favorable despite the loss in statistical power. In the original analysis of this model, using a sample size of 1000 with no replication, Segovia-Juarez et al. (2004) find nine significant parameters, including parameters α_{BI} and λ. Reducing sample size to 300 with no replication (Fig. 9, black bars), for a total of 300 model simulations, parameters α_{BI} and λ are not significantly different from zero. Using 100 samples and averaging three replicates (Fig. 9, gray bars), for a total of 300 model simulations, however, successfully identifies α_{BI} and λ as significant, yielding performance similar to that of 1000 samples used by Segovia-Juarez et al. Therefore, for this ABM we suggest that performing replicate simulations of each parameter sample and averaging these replicates is a useful strategy that can provide better performance with fewer computationally intensive model simulations.

5.5. eFAST results: aleatory uncertainty mishandled in total-order STi

To our knowledge, eFAST has been used to analyze an ABM once before (Lempert et al., 2002), but details of the methodology used are lacking. To explore the suitability of the eFAST method in analysis of stochastic models, we analyze the Segovia-Juarez model, described above, taking advantage of the replication and averaging scheme (described in the previous section) and of the use of a dummy parameter (as developed in Section 3.3.1). eFAST is performed using 257 samples per search curve (i.e. $N_S = 257$, chosen to satisfy condition (A.9)), a resampling size of 4 (i.e., $N_R = 4$), and four replicates for averaging. By allowing eFAST to partition variance to a dummy parameter, we find that the first-order index is relatively unaffected (Fig. 10A). However, eFAST severely mishandles the variability induced by aleatory uncertainty by incorrectly partitioning it to the total-order sensitivity index, resulting in a large artifactual $S_{Ti} = 24.3\%$ (Fig. 10B, dummy parameter, black bar). Thus, use of a dummy parameter to quantify this artifact reveals that parameters with a total-order S_{Ti} in the range of ~20–30% are likely artifactual. Using replication and averaging to limit variability induced by aleatory uncertainty successfully reduces this artifact by more than half (Fig. 10B, dummy parameter, gray bar): this allows seven of the eight significant parameters identified by LHS/PRCC to exceed the dummy parameter background value for S_i (Fig. 10A). Increasing N_S and N_R, as well as performing averaging with additional replicates will likely further improve the limit of detection. Performing significance testing in combination with replication and averaging, we find that eight of the nine parameters identified as significant by LHS/PRCC are also identified by eFAST first-order S_i (Fig. 10A), while the S_{Ti} coefficients are significantly different from the dummy only for four of the eight parameters listed by LHS/PRCC (Fig. 10B). Note that eFAST results come at a much higher computational cost: 53,456 total model simulations using eFAST, as opposed to 300 total model simulations with LHS/PRCC.

Summary: This ABM example shows how LHS/PRCC and eFAST are typically in agreement, identifying a similar set of important parameters (with the eFAST sets of important parameters being smaller than PRCC, as shown in some of the deterministic model examples in Section 4). However, these two methods have different relative strengths and weaknesses. eFAST requires many more model simulations, which is a particular problem as ABMs are often more computationally intensive than deterministic models. Also, when analyzing stochastic models, eFAST produces an artifact whereby aleatory variance is inappropriately partitioned to the total-order sensitivity index. Therefore, this example suggests how the LHS/PRCC method (with the modifications presented here) can perform better than eFAST for a specific time point, reaching the same conclusions with much less computational cost.

The eFAST method has the strength of identifying non-monotonic sensitivities, however. Though we see no evidence of this strength in the ABM we analyze here, there is no *a priori* way of knowing if non-monotonic sensitivities are present in a model. Therefore, if computational cost is not prohibitive, use of eFAST on stochastic models can complement LHS/PRCC results. In this case, one should take steps to reduce the artifact of eFAST mishandling aleatory variance in the total-order sensitivity index, or one should rely solely on the first-order index, as it is unaffected by the artifact.

6. Discussion and conclusion

Uncertainty and sensitivity (US) analyses offer a way to assess the adequacy of models and establish what factors affect model

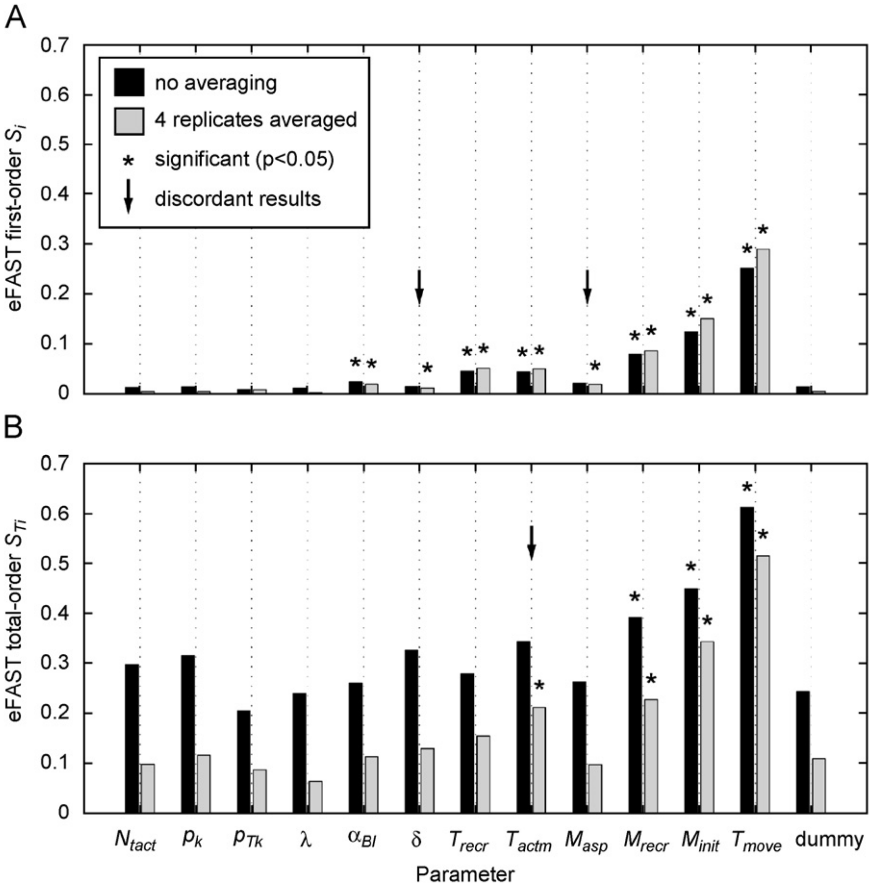

Fig. 10. Performance of eFAST on the stochastic agent-based model of Segovia-Juarez et al. (2004). (A) eFAST first-order sensitivity index S_i. Replication and averaging (gray bars) have a small effect on S_i, allowing additional parameters (indicated by arrow) to be identified as statistically significant (*).(B) eFAST total-order sensitivity index S_{Ti}. eFAST artifactually assigns aleatory variance to the total-order index, resulting in a high dummy parameter background value (dummy parameter, black bar). Replication and averaging (gray bars) partially reduce this artifact, allowing an additional parameter (indicated by arrow) to be found significant (*).

outputs. We reviewed and compared two specific types of global sensitivity analysis (SA) indexes that have proven to be among the most reliable and efficient, namely a sampling-based method (partial rank correlation coefficient—PRCC) and a variance-based method (extended Fourier amplitude sensitivity test—eFAST). All functions used throughout the paper are available on our website (http://malthus.micro.med.umich.edu/lab/usanalysis.html).

PRCC provides answers to questions such as how the output is affected if we increase (or decrease) a specific parameter (linearly discounting the effects of the uncertainty over the rest of the parameters). Thus PRCC can be informative on what parameters to target if we want to achieve specific goals (e.g. control or regulatory mechanisms). For example, the most significant set of parameters can be used to determine how to efficiently reduce viral load or increase immune response (by both timing and magnitude). eFAST, and all variance-based methods in general, indicate which parameter uncertainty has the greatest impact on output variability. In other words, our predictions will be strengthened if we can reduce uncertainty and get better estimates on specific parameters of the model (i.e. the ones with highest S_i and S_{Ti} sensitivity coefficients). This will also enhance any additional PRCC or sampling-based analysis, because any regulatory or control strategy will be more reliable.

A general finding was that a properly designed US analysis returns a set of bifurcation parameters as significant; thus US analysis can be an adequate alternative when an analytical solution to a mathematical system is not possible. This holds true for both sample and variance-based methods (at least if we consider the examples of Section 4).

One critical point is the selection of adequate (pdfs) and the choice of parameter ranges used for sampling: The selection of probability distributions for the uncertain parameters depends in part on whether the intent of the analysis is an exploration of variable effects (i.e., a SA) or a propagation of uncertainty to assess the uncertainty in the outcomes of interest (i.e., an uncertainty analysis (US)). For meaningful US, the selected distributions are chosen based on the degree of our understanding of biology with respect to the appropriate values of model parameters. In contrast, the distributions might be selected simply to fully explore potential variable effects in a SA.

Choice of parameter ranges should also be guided by the available knowledge of the biological problem. If a parameter range is completely outside the biological realm, the practical relevance of the US analysis is lost. Unfortunately, exact or even hypothetical biological ranges are many times unknown. Unless the choice of very small ranges for certain parameters is guided by some *a priori* knowledge, the sampling should be performed within the whole range of plausible values, since US analysis results can be quite different (see Supplement F, Fig F.1 and Table F.1 for examples of how PRCC and eFAST results can be affected by sampling different ranges). Moreover, if the range of variation for some parameter spans several orders of magnitude and we want to avoid under-sampling in the outer ranges of parameter space, the implementation of a logarithmic sampling scheme (for example, a log-uniform in place of a uniform pdf) would achieve the goal, but the final results will be affected (as shown in Supplement D online, Tables D.1 and D.3).

Our examples showed how PRCC and eFAST could give different results, where nonlinear and non-monotonic relationships between inputs and output lead to misleading conclusions, even with a simple prey–predator model (see Section 4.1). eFAST is more general than PRCC (it deals with any type of relationship between inputs and outputs of a model) with one major drawback: its computational cost, especially for computing S_{Ti}. Our examples highlight how a large number of iterations are needed to achieve a recommended degree of accuracy in eFAST (as also shown in Fig. 8.9 at p. 190 of Saltelli et al., 2000).

We designed a new method (together with a heuristic for its accuracy) to check for significant eFAST S_i and S_{Ti}. This new method is based on a two-sample *t*-test approach and uses a resampling scheme to compare each sensitivity index with that of a dummy parameter. The use of dummy parameters is a standard practice in screening methods (see Saltelli et al., 2000, Chapter 4), although, to our knowledge, it has never been applied in the context of eFAST.

We built several functions to display scatter plots of sampled parameters versus the output under study (given on our website, see http://malthus.micro.med.umich.edu/lab/usanalysis.html): they can be extremely useful to detect non-monotonicities and nonlinearities (as shown in Section 4) and possibly explain why PRCC and eFAST results are different.

In the context of agent-based model (ABM), we also implemented a new averaging strategy to efficiently and reliably perform US analysis within the context of this particular class of stochastic models. By running the ABM multiple times with the same set of parameters and then averaging the output, we are able to attenuate the effect of aleatory uncertainty, obtaining more reliable results both for PRCC and eFAST.

We also implemented a concordance measure (top-down coefficient of concordance—TDCC) of the adequacy of the sample size and can check if PRCC and eFAST return a similar set of important parameters.

Throughout our examples, we found that the set of significant first-order S_i and total-order S_{Ti} sensitivity indexes returned by eFAST is consistent and with the same ranking. eFAST generally returns smaller subsets of parameters with significant S_i and S_{Ti}, compared to the set of parameters with significant PRCC. Also, eFAST often returns S_i and S_{Ti} for parameters that are not listed as significant by PRCC and, when a similar set of parameters is returned, the order of importance is frequently different between the two methods.

Based on the examples shown, we conclude that PRCC and eFAST S_i and S_{Ti} should both be computed, keeping in mind that they measure different things, and that eFAST is generally computationally more expensive. If the computational cost is too high (i.e. cpu time per one model run), more efficient methods are available (e.g. methods that are less affected by non-monotonicities). A lower computational cost can be achieved by reducing N_S and N_R (i.e. the total number of runs), but the accuracy of the results might be affected. To increase the accuracy of eFAST results, in general, it is better to increase N_S first (and check for the TDCC to see if that sample size is sufficient to capture a similar set of most important parameters), and then eventually increase N_R (to improve the accuracy of inference on the indexes). We found that values and significances of first-order S_i are affected by increasing N_S: the values of the not significant S_i get smaller, as does the dummy parameter S_{dummy}, making the *t*-test more reliable. Increasing N_S does not affect S_{Ti} values, although the *t*-test becomes more precise.

In general, the computational execution time of the model is the major concern when performing US analysis (although it was not a major issue for the examples given in Section 4 and for the ABM). Models with several parameters and many complex nonlinear mechanisms likely result in high computational costs. Screening methods are available to address this problem. Within the class of screening methods, Morris (1991) (or elementary effects) is the most popular: it is global, computationally efficient, and should be implemented as a first preliminary US analysis when the execution time of the model is prohibitive (several hours or days).

In summary, characterizing uncertainty in parameter values and initial conditions in mathematical models in biology is attainable and is dependent on the type of system under study. For the math biology community to continue to gain a foothold and have an impact in important biology problems, it is clear that identifying uncertainty in our models is of key importance. By the very act of classifying this uncertainty, we can simultaneously identify the parameters (i.e. biological mechanisms) that are driving system outputs. These mechanisms can then be posed to the experimental community to test. This close interaction between theorists and experimentalists provides the greatest opportunity for the use of mathematical models.

Acknowledgments

This work was supported by NIH Grants HL68526, LM 009027, and DAIT-BAA-05-10. We are grateful to Jennifer Linderman, Rick Riolo, and to the members of the Kirschner laboratory for helpful discussions. Also, comments by two anonymous reviewers were invaluable in making this manuscript as comprehensive and clear as possible.

Appendix A. Supplementary Materials

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.jtbi.2008.04.011.

References

Anderson, T.W., 2003. An Introduction to Multivariate Statistical Analysis, third ed. Wiley Series in Probability and Statistics. Wiley-Interscience, Hoboken, NJ.

Apostolakis, G., 1990. The concept of probability in safety assessments of technological systems. Science 250 (4986), 1359–1364.

Blower, S.M., Dowlatabadi, H., 1994. Sensitivity and uncertainty analysis of complex-models of disease transmission—an HIV model, as an example. Int. Stat. Rev. 62 (2), 229–243.

Cacuci, D.G., Ionescu-Bujor, M., 2004. A comparative review of sensitivity and uncertainty analysis of large-scale systems—II: statistical methods. Nucl. Sci. Eng. 147 (3), 204–217.

Chang, S.T., Linderman, J.J., Kirschner, D.E., 2008. Multiple polymorphisms on antigen presentation and susceptilbity to *M. tuberculosis* infection. Infect. Immun., doi:10.1128/IAI.01677-07.

Collins, D.C., Avissar, R., 1994. An evaluation with the Fourier amplitude sensitivity test (FAST) of which land-surface parameters are of greatest importance in atmospheric modeling. J. Climate 7 (5), 681–703.

Cooke, R., 1991. Experts in Uncertainty: Opinion and Subjective Probability in Science. Environmental Ethics and Science Policy. Oxford University Press, New York.

Cukier, R.I., Fortuin, C.M., Shuler, K.E., Petschek, A.G., Schaibly, J.H., 1973. Study of sensitivity of coupled reaction systems to uncertainties in rate coefficients 1. Theory. J. Chem. Phys. 59 (8), 3873–3878.

Draper, D., 1995. Assessment and propagation of model uncertainty. J. R. Stat. Soc. B—Methodological 57 (1), 45–97.

Evans, J.S., Gray, G.M., Sielken, R.L., Smith, A.E., Valdezflores, C., Graham, J.D., 1994. Use of probabilistic expert judgment in uncertainty analysis of carcinogenic potency. Regul. Toxicol. Pharmacol. 20 (1), 15–36.

Helton, J.C., 1993. Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive-waste disposal. Reliab. Eng. Syst. Saf. 42 (2–3), 327–367.

Helton, J.C., 1997. Uncertainty and sensitivity analysis in the presence of stochastic and subjective uncertainty. J. Stat. Comput. Simulation 57 (1–4), 3–76.

Helton, J.C., 1999. Uncertainty and sensitivity analysis in performance assessment for the waste isolation pilot plant. Comput. Phys. Commun. 117 (1–2), 156–180.

Helton, J.C., Breeding, R.J., 1993. Calculation of reactor accident safety goals. Reliab. Eng. Syst. Saf. 39 (2), 129–158.

Helton, J.C., Davis, F.J., 2002. Illustration of sampling-based methods for uncertainty and sensitivity analysis. Risk Anal. 22 (3), 591–622.

Helton, J.C., Davis, F.J., 2003. Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems. Reliab. Eng. Syst. Saf. 81 (1), 23–69.

Helton, J.C., Johnson, J.D., Shiver, A.W., Sprung, J.L., 1995. Uncertainty and sensitivity analysis of early exposure results with the Maccs reactor accident consequence model. Reliab. Eng. Syst. Saf. 48 (2), 91–127.

Helton, J.C., Martell, M.A., Tierney, M.S., 2000. Characterization of subjective uncertainty in the 1996 performance assessment for the waste isolation pilot plant. Reliab. Eng. Syst. Saf. 69 (1–3), 191–204.

Helton, J.C., Johnson, J.D., Sallaberry, C.J., Storlie, C.B., 2006. Survey of sampling-based methods for uncertainty and sensitivity analysis. Reliab. Eng. Syst. Saf. 91 (10–11), 1175–1209.

Helton, J.C., Johnson, J.D., Oberkampf, W.L., Storlie, C.B., 2007. A sampling-based computational strategy for the representation of epistemic uncertainty in model predictions with evidence theory. Comput. Methods Appl. Mech. Eng. 196 (37–40), 3980–3998.

Hoare, A., Regan, D.G., Wilson, D.P., 2008. Sampling and sensitivity analyses tools (SaSat) for computational modeling. Theor. Biol. Med. Model. 5, 4 (open access at 〈http://www.tbiomed.com/content/5/1/4〉).

Hora, S.C., Helton, J.C., 2003. A distribution-free test for the relationship between model input and output when using Latin hypercube sampling. Reliab. Eng. Syst. Saf. 79 (3), 333–339.

Hora, S.C., Iman, R.L., 1989. Expert opinion in risk analysis—the Nureg-1150 methodology. Nucl. Sci. Eng. 102 (4), 323–331.

Iman, R.L., Conover, W.J., 1980. Small sample sensitivity analysis techniques for computer-models, with an application to risk assessment. Commun. Stat. A—Theory Methods 9 (17), 1749–1842.

Iman, R.L., Conover, W.J., 1982. A distribution-free approach to inducing rank correlation among input variables. Commun. Stat. B—Simulation Comput. 11 (3), 311–334.

Iman, R.L., Conover, W.J., 1987. A measure of top-down correlation. Technometrics 29 (3), 351–357.

Iman, R.L., Davenport, J.M., 1982. Rank correlation plots for use with correlated input variables. Commun. Stat. B—Simulation Comput. 11 (3), 335–360.

Iman, R.L., Helton, J.C., 1988. An investigation of uncertainty and sensitivity analysis techniques for computer models. Risk Anal. 8 (1), 71–90.

Kirschner, D.E., Chang, S.T., Riggs, T.W., Perry, N., Linderman, J.J., 2007. Toward a multiscale model of antigen presentation in immunity. Immunol. Rev. 216, 93–118.

Kleijnen, J.P.C., Helton, J.C., 1999. Statistical analyses of scatterplots to identify important factors in large-scale simulations, 1: review and comparison of techniques. Reliab. Eng. Syst. Saf. 65 (2), 147–185.

Lempert, R., Popper, S., Bankes, S., 2002. Confronting surprise. Soc. Sci. Comput. Rev. 20 (4), 420–440.

Lotka, A.J., 1925. Elements of Physical Biology. Williams & Wilkins Co., Baltimore.

Marino, S., Kirschner, D.E., 2004. The human immune response to *Mycobacterium tuberculosis* in lung and lymph node. J. Theor. Biol. 227 (4), 463–486.

Marino, S., Pawar, S., Fuller, C.L., Reinhart, T.A., Flynn, J.L., Kirschner, D.E., 2004. Dendritic cell trafficking and antigen presentation in the human immune response to *Mycobacterium tuberculosis*. J. Immunol. 173 (1), 494–506.

Marino, S., Beretta, E., Kirschner, D.E., 2007. The role of delays in innate and adaptive immunity to intracellular bacteria infection. Math. Biosci. Eng. 4 (2), 261–286.

McKay, M., Meyer, M., 2000. Critique of and limitations on the use of expert judgements in accident consequence uncertainty analysis. Radiat. Prot. Dosim. 90 (3), 325–330.

Mckay, M.D., Beckman, R.J., Conover, W.J., 1979. Comparison of 3 methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21 (2), 239–245.

Morris, M.D., 1991. Factorial sampling plans for preliminary computational experiments. Technometrics 33 (2), 161–174.

Morris, M.D., 2000. Three technometrics experimental design classics. Technometrics 42 (1), 26–27.

Parry, G.W., Winter, P.W., 1981. Characterization and evaluation of uncertainty in probabilistic risk analysis. Nucl. Saf. 22 (1), 28–42.

Pate'-Cornell, M.E., 1996. Uncertainties in risk analysis: six levels of treatment. Reliab. Eng. Syst. Saf. 54 (2–3), 95–111.

Perelson, A.S., Kirschner, D.E., De Boer, R., 1993. Dynamics of HIV infection of CD4+ T cells. Math. Biosci. 114 (1), 81–125.

Ratto, M., Pagano, A., Young, P., 2007. State dependent parameter metamodelling and sensitivity analysis. Comput. Phys. Commun. 177 (11), 863–876.

Riggs, T., Walts, A., Perry, N., Bickle, L., Lynch, J.N., Myers, A., Flynn, J., Linderman, J.J., Miller, M.J., Kirschner, D.E., 2008. A comparison of random vs. chemotaxis-driven contacts of T cells with dendritic cells during repertoire scanning. J. Theor. Biol. 250 (4), 732–751.

Saltelli, A., 2002. Making best use of model evaluations to compute sensitivity indices. Comput. Phys. Commun. 145 (2), 280–297.

Saltelli, A., 2004. Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models. Wiley, Hoboken, NJ.

Saltelli, A., Bolado, R., 1998. An alternative way to compute Fourier amplitude sensitivity test (FAST). Comput. Stat. Data Anal. 26 (4), 445–460.

Saltelli, A., Marivoet, J., 1990. Nonparametric statistics in sensitivity analysis for model output—a comparison of selected techniques. Reliab. Eng. Syst. Saf. 28 (2), 229–253.

Saltelli, A., Tarantola, S., Chan, K.P.S., 1999. A quantitative model-independent method for global sensitivity analysis of model output. Technometrics 41 (1), 39–56.

Saltelli, A., Chan, K., Scott, E.M., 2000. Sensitivity Analysis. Wiley Series in Probability and Statistics. Wiley, Chichester, New York.

Saltelli, A., Ratto, M., Tarantola, S., Campolongo, F., 2005. Sensitivity analysis for chemical models. Chem. Rev. 105 (7), 2811–2827.

Savage, I.R., 1956. Contributions to the theory of rank order-statistics—the 2-sample case. Ann. Math. Stat. 27 (3), 590–615.

Schaibly, J.H., Shuler, K.E., 1973. Study of sensitivity of coupled reaction systems to uncertainties in rate coefficients 2. Applications. J. Chem. Phys. 59 (8), 3879–3888.

Segovia-Juarez, J.L., Ganguli, S., Kirschner, D., 2004. Identifying control mechanisms of granuloma formation during *M. tuberculosis* infection using an agent-based model. J. Theor. Biol. 231 (3), 357–376.

SimLab. 2006. Distributed under the SimLab Software License, Version 1.0 〈http://simlab.jrc.ec.europa.eu/〉.

Storlie, C.B., Helton, J.C., 2008a. Multiple predictor smoothing methods for sensitivity analysis: description of techniques. Reliab. Eng. Syst. Saf. 93 (1), 28–54.

Storlie, C.B., Helton, J.C., 2008b. Multiple predictor smoothing methods for sensitivity analysis: example results. Reliab. Eng. Syst. Saf. 93 (1), 55–77.

Tarantola, S., Gatelli, D., Mara, T.A., 2006. Random balance designs for the estimation of first order global sensitivity indices. Reliab. Eng. Syst. Saf. 91 (6), 717–727.

Volterra, V., 1926. Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. Mem. R. Accad. Naz. dei Lincei. Ser. VI, vol. 2 (English version can be found in Animal Ecology. McGraw-Hill, 1931. Translated from 1926 edition by R.N. Chapman. Title "Variations and fluctuations of the number of individuals in animal species living together.").