Statistica Sinica Preprint No: SS-2016-0403R1	
Title	Bayesian Calibration of Multistate Stochastic Simulators
Manuscript ID	SS-2016-0403R1
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0403
Complete List of Authors	Matthew Pratola and Oksana Chkrebtii,
Corresponding Author	Matthew Pratola
E-mail	hbindele@southalabama.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

Statistica Sinica

BAYESIAN CALIBRATION OF MULTISTATE STOCHASTIC SIMULATORS

M. T. Pratola and O. Chkrebtii

Department of Statistics, The Ohio State University

Abstract:

Inference on large-scale models is of great interest in modern science. Examples include deterministic simulators of fluid dynamics to recover the source of a pollutant, and stochastic agent-based simulators to infer features of consumer behaviour. When computational constraints prohibit model evaluation at all but a small ensemble of parameter settings, exact inference is infeasible. In such cases, emulation of the simulator enables the interrogation of a surrogate model at arbitrary parameter values. Combining emulators with observational data to estimate parameters and predict a real-world process is known as computer model calibration. The choice of the emulator model is a critical aspect of calibration. Existing approaches treat the mathematical model as implemented on computer as an unknown but deterministic response surface. In many cases the underlying mathematical model, or the simulator approximating the mathematical model, are not determinsitic and in fact have some uncertainty associated with their output. In this paper, we propose a Bayesian statistical calibration model for stochastic simulators. The approach is motivated by two applied problems: a deterministic mathematical model of intra-cellular signalling whose implementation on computer nonetheless has discretization uncertainty, and a stochastic model of river

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

water temperature commonly used in hydrology. We show the proposed approach is able to map the uncertainties of such non-deterministic simulators through to the resulting inference while retaining computational feasibility. Supplementary computer code and datasets are provided online.

Key words and phrases: Computer Experiments, Uncertainty Quantification, Differential Equation, Stochastic Simulation, Physical Statistical, Models

1. Introduction

Models of complex processes allow scientists to gain a deeper understanding of system dynamics or enable policy makers to make decisions based on future projections. These models, known as computer simulators, may solve large-scale systems of differential equations or implement stochastic simulations such as agent-based systems, that describe real-world processes. Of particular importance to decision makers is the task of appropriately quantifying and combining uncertainty from all sources when performing inference.

More specifically, simulators can be said to describe the spatio-temporal evolution of one or many system states, defined up to some unknown components called calibration parameters. These may include physical constants, forcing functions, or initial or boundary conditions. For a given computer model, interest lies in inferring unknown calibration parameters from noisy, often indirect observations of the states at discrete spatio-temporal locations. An important challenge arises when

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

the states, and hence the likelihood of the data, are computationally expensive to evaluate. Computer model calibration (Kennedy and O'Hagan (2001); Higdon et al. (2004, 2008); Goldstein and Rougier (2006); Joseph and Melkote (2009)) performs inference in this situation by modeling, or emulating, the simulated states conditional on a well-designed sample of the computationally expensive simulator. The additional source of uncertainty associated with the emulation is propagated through the inference, typically using a hierarchical Bayesian framework. Our work in this paper is concerned with accounting for stochasticity in the state, a key source of uncertainty that has so far been mostly ignored or, at best, inadequately represented in the statistical calibration literature.

Existing methodology essentially treats simulators as deterministic black-box functions, where the output is fixed for a given parameter input setting. That is, it is assumed that running the simulator at the same inputs will always produce exactly the same output. However, it is widely known that in a broad class of problems this assumption is unrealistic, and a given parameter input setting will yield a sample of realizations, or ensembles, from an unknown distribution over the states.

For instance, agent-based models aim to reconstruct the macroscopic behaviour of complex systems by forward simulating a large number of "agent" models that describe the microscopic behaviour of the system under study. Such models are used in analyzing the behaviour of the stock market and biological systems (Palmer

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

et al. (1994); Tesfatsion (2002); Gilbert (2008); Auchincloss and Roux (2008)). These stochastic simulation models also make available an ensemble of solution realizations at given settings of the parameters. We investigate a stochastic simulator of water temperature (Cluis (1972); Caissie et al. (1998)), where river water temperature is simulated by combining sparse observational water data with readily available air temperature data. The goal is to calibrate scientifically meaningful air-to-water heat transfer coefficients.

Stochastic simulations also arise when, for a given input, the output states are deterministic but uncertain. For example, a simulator defined implicitly as a set of partial differential equations (PDE) typically does not have a closed form solution. Instead, for a given parameter setting, the states are discretized and approximated numerically using a deterministic technique. It has been shown that choices related to this discretization can have a substantial effect on approximated system states (e.g. Kim et al. (2013); Arridge et al. (2006)), so that a typical calibration framework that ignores this error is likely to lead to biased estimates of the calibration parameters and posterior under-coverage. This issue has led to the use of Bayesian ideas for modeling uncertainty associated with discretization of an infinite-dimensional state as a stochastic process (Chkrebtii et al. (2016)). However, as with discretizing the PDE system, simulating realizations from this probabilistic uncertainty model is typically computationally expensive. Instead, an ensemble of solution realizations

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

of the probabilistic solver of Chkrebtii et al. (2016) may be obtained at a small, well-chosen collection of calibration parameter settings and used to perform efficient inference in the approach we propose.

One example of an implicit model where the solution states have non-negligible discretization uncertainty describes the temporal evolution of the concentration of four intracellular gene transcription factors within the JAK-STAT signalling network pathway (Pellegrini and Dusanter-Fourt (1997); Swameye et al. (2003)). It has been shown in Chkrebtii et al. (2016) that choices related to discretization strongly shape posterior correlations among model states, motivating the use of simulators that model this uncertainty. Not only are such simulators stochastic, but they are computationally expensive, motivating further advances in computer model calibration.

This paper is concerned with developing a statistical approach to computer model calibration experiments which can take into account the uncertainty in simulation models when made available as a large ensemble of realizations. Our approach uses empirical orthogonal functions to represent the functional uncertainty of the simulator by associating each ensemble member realized at a given setting of the calibration parameter with a single latent weight. These latent weights are then modeled as points in a latent weight-space on which we place a Gaussian process prior which we can then use to construct unobserved realizations of the simulation model at unobserved settings of the parameters while retaining the desired uncertainty.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

The reconstructed simulator realization at the unknown parameter corresponding to the observational data is linked through a hierarchical Bayesian model for the field observations. Included in this observational data model are model discrepancy components that are also given Gaussian process priors. The overall model specification is then completed by placing appropriate prior distributions on model parameters, and the model is fitted by a Markov chain Monte Carlo (MCMC) algorithm.

We first begin by reviewing the concept of calibration for computer experiments in the context of Bayesian hierarchical modeling. We then describe the Kennedy-O'Hagan model that forms the basis of further developments.

1.1 Calibration Experiments

The problem of inference, or calibration, for computer models of a state $x(\mathbf{s};\boldsymbol{\theta})$ at spatial-temporal locations $\mathbf{s}_i \in \mathcal{S}$ and unknown calibration parameter setting $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ consists of recovering the unknown calibration parameters $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ from partial or indirect observations, $y(\mathbf{s})$, of the state. The calibration parameters $\boldsymbol{\theta}$ represent the setting of this parameter that "best" matches the computer model to the observed data. They usually are themselves of considerable scientific interest when these parameters have important scientific meaning, such as the viscosity of a modeled fluid or the initial state of a dynamical system.

Because the simulator is often an inexact representation of reality, the notion of a systemic discrepancy is introduced between the simulator and the true state of

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

the observed process. Such discrepancy may be additive, represented by $\delta(\mathbf{s})$, which allows for correcting an additive bias in the simulated state as $x(\mathbf{s};\boldsymbol{\theta})+\delta(\mathbf{s})$. Another popular correction is multiplicative discrepancy, represented as κ and usually taken to be constant with respect to spatial-temporal location. This discrepancy allows for correcting the scaling of the simulated state as $\kappa x(\mathbf{s};\boldsymbol{\theta})$.

Let $\mathbf{x}(\boldsymbol{\theta}) = (x(\mathbf{s}_1;\boldsymbol{\theta}),\ldots,x(\mathbf{s}_n;\boldsymbol{\theta}))^\top$ represent the vector of state outputs at the spatial-temporal grid locations $\mathbf{s}_1,\ldots,\mathbf{s}_n$, let $\boldsymbol{\delta} = (\delta(\mathbf{s}_1),\ldots,\delta(\mathbf{s}_n))^\top$ represent the vector of the additive discrepancy at spatial-temporal grid locations $\mathbf{s}_1,\ldots,\mathbf{s}_n$ and let $\boldsymbol{\Lambda}_f$ be an $n\times n$ precision matrix representing the uncertainty in our observations. Then the likelihood of the observations, $\mathbf{y} = (y(\mathbf{s}_1),\ldots,y(\mathbf{s}_n))^\top$, observed at the n spatial-temporal locations conditional on the state outputs, calibration parameters, discrepancies and precision parameters is

$$\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\delta}, \kappa, \boldsymbol{\Lambda}_f.$$

The simplest model (Higdon et al. (2004)) assumes homoscedastic precision, $\boldsymbol{\Lambda}_f = \lambda_f \mathbf{I}_n$ and a Gaussian likelihood, so that the conditional distribution is

$$\mathbf{y} \mid \boldsymbol{\theta}, \boldsymbol{\delta}, \kappa, \lambda_f \sim N\left(\kappa\mathbf{x}(\boldsymbol{\theta}) + \boldsymbol{\delta}, \lambda_f\mathbf{I}\right). \tag{1.1}$$

If the simulator were computationally inexpensive, estimating the unknowns would be fairly straightforward – specifying priors on the calibration parameter, $\pi(\boldsymbol{\theta})$, discrepancies, $\pi(\boldsymbol{\delta},\kappa)$ and precision, $\pi(\lambda_f)$, one could sample from the poste-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

rior distribution

$$\boldsymbol{\theta}, \boldsymbol{\delta}, \kappa, \lambda_f \mid \mathbf{y}$$

using a Metropolis within Gibbs algorithm (Higdon et al. (2004)), which requires evaluating the simulator at a large number of proposed settings of the calibration parameter, $\boldsymbol{\theta}$.

However, due to the high computational cost of producing simulations of the state $\mathbf{x}(\boldsymbol{\theta})$, only a limited number, say m, of simulator evaluations, can be made. This feature of the simulator immediately precludes the use of any inferential approach which requires large numbers of simulator evaluations at settings of $\boldsymbol{\theta}$, such as the approach just described.

This computational limitation led to the introduction of an additional layer in the Bayesian hierarchy representing uncertainty in the simulator $\mathbf{x}(\boldsymbol{\theta})$, which is *emulated* rather than being evaluated. The emulator is a statistical model for the state given a small well-designed collection of m simulator evaluations, $\mathbf{x}(\boldsymbol{\theta}_1), \ldots, \mathbf{x}(\boldsymbol{\theta}_m)$. This conditional distribution of the state at the calibration parameter setting $\boldsymbol{\theta}$ given the m state outputs evaluated at parameter settings $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m$ is expressed as

$$\mathbf{x}(\boldsymbol{\theta}) \mid \mathbf{x}(\boldsymbol{\theta}_1), \ldots, \mathbf{x}(\boldsymbol{\theta}_m), \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m, \boldsymbol{\theta}, \cdot.$$

It may depend on additional hyperparameters (here denoted by the '$\cdot$'), the form of which depends on the specific emulation model used. For instance, Kennedy

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

and O'Hagan (2001) use a Gaussian process (GP, Sacks et al. (1989)) emulator, while Higdon et al. (2008) use basis functions for dimension reduction in addition to a Gaussian process model. In any case, the introduction of this second layer of modeling allows one to construct predictions of the unobserved state for arbitrary choices of calibration setting $\boldsymbol{\theta}$ as well as propagating the uncertainty in the emulated state through to the posterior inference for the calibration parameter and all other quantities of interest.

1.2 The Kennedy-O'Hagan Model

The method proposed by Kennedy and O'Hagan (2001) is widely considered as the basis for subsequent development of statistical computer model calibration, so we elaborate on it in relation to our general setup. Their approach, subsequently expanded into a fully Bayesian approach by Higdon et al. (2004, 2008), makes extensive use of Gaussian process (GP) priors and Gaussian conjugacy. The likelihood for the observations is specified as (1.1), while the state is modeled a priori as a realization of a GP,

$$\begin{pmatrix} \mathbf{x}(\boldsymbol{\theta}) \\ \mathbf{x} \end{pmatrix} \sim N\left(\boldsymbol{\mu}, \lambda_x^{-1} \begin{bmatrix} \mathbf{R}_0 & \mathbf{R}_{0,\mathbf{x}} \\ \mathbf{R}_{\mathbf{x},0} & \mathbf{R}_{\mathbf{x}} \end{bmatrix} + \lambda_c^{-1}\mathbf{I}_{n(m+1)}\right),$$

where $\mathbf{x} = \left(\mathbf{x}(\boldsymbol{\theta}_1)^T, \ldots, \mathbf{x}(\boldsymbol{\theta}_m)^T\right)^T$, $\boldsymbol{\mu} = \left(\boldsymbol{\mu}_0^T, \boldsymbol{\mu}_x^T\right)^T \in \mathbb{R}^{(m+1)n}$ is the mean of the states, $\lambda_x^{-1} \in \mathbb{R}$ is the marginal process variance, and $\lambda_c^{-1} \in \mathbb{R}$ represents small scale variability of the states, sometimes called the "nugget" (Cressie (1993)). The

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

correlation matrix is typically modeled using the Gaussian correlation function, that assumes the states can be represented by a smooth, infinitely differentiable process, and parameterized as

$$[\mathbf{R}_{\mathbf{x}}]_{ij} = \prod_{k=1}^{p}\prod_{l=1}^{q} \phi_k^{(s_{ik}-s_{jk})^2} \rho_l^{(\theta_{il}-\theta_{jl})^2},$$

where the $\phi_k \in (0,1)$ are correlation parameters for all $k = 1, \ldots, p$ spatial-temporal covariate dimensions and the $\rho_l \in (0,1)$ are correlation parameters for all $l = 1, \ldots, q$ calibration parameter dimensions.

The discrepancy is also modeled as a realization of a GP,

$$\boldsymbol{\delta} \sim N(\boldsymbol{\mu}_\delta, \lambda_\delta^{-1}\mathbf{R}_\delta),$$

where $\boldsymbol{\mu}_\delta \in \mathbb{R}^n$, $\lambda_\delta \in \mathbb{R}$, and $[\mathbf{R}_\delta]_{ij} = \prod_{k=1}^{p} \psi_k^{(s_{ik}-s_{jk})^2}$, which models a smooth discrepancy between the calibrated simulator and the observed process with correlation parameters $\psi_k \in (0,1), \quad k = 1, \ldots, p$.

Combining these priors with the likelihood, the joint model of Kennedy and O'Hagan (2001) for the field observations and simulator outputs is,

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_0 + \boldsymbol{\mu}_\delta \\ \boldsymbol{\mu}_x \end{pmatrix}, \lambda_x^{-1}\begin{bmatrix} \mathbf{R}_0 & \mathbf{R}_{0,x} \\ \mathbf{R}_{x,0} & \mathbf{R}_x \end{bmatrix} + \lambda_\delta^{-1}\begin{bmatrix} \mathbf{R}_\delta & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} \lambda_f^{-1}\mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & \lambda_c^{-1}\mathbf{I}_{nm} \end{bmatrix}\right). \tag{1.2}$$

This model has been discussed at length in the computer experiments literature. There are assumptions of it that do not satisfy our requirements. The term λ_c^{-1},

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

which represents simulator output uncertainty, has largely been dealt with in a cursory manner or simply ignored. Primarily, the setting of this parameter has been driven by a desire to maintain computational stability in manipulating the large covariance matrices of (1.2) rather than a concerted attempt to model and quantify possible uncertainties in simulator outputs. Furthermore, a simple i.i.d. Normal error model is likely not justified to account for simulator error as most of the simulation models calibrated in this framework exhibit smooth and continuously varying response surfaces as functions of the simulator's inputs. More realistic uncertainty is sometimes available when an ensemble of realizations of a stochastic simulation model are available. Model (1.2) was proposed in the context of calibrating a single state. In many applications, one may be interested in calibrating multiple states simultaneously, some or all of which are observed in the field. Extending (1.2) to the case of multiple states would seem difficult given the computational limitations of the model with just a single state.

In the next section, we motivate the need for a statistical calibration methodology that can account for simulator uncertainties, and potentially multiple states, with an application in water temperature modeling and a PDE model of a biochemical system. We develop our model in Section 3, and demonstrate the proposed approach on the water temperature and JAK-STAT examples in Sections 4 and 5. We conclude in Section 6.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

2. Motivation

In this section, we introduce two examples of calibrating simulators to observations where simulator uncertainty need be accounted for in the statistical methodology. We are also interested in calibrating multi-state stochastic simulators, and our second example involves calibrating four states.

2.1 Stochastic Water Temperature Model

The prediction of temperature fluctuations in inland bodies of water is critical for ecological and conservation initiatives because of its effect on wildlife and the possibility of monitoring thermal water pollution. Climate change has made such studies increasingly important in order to understand and predict water quality and aquasystem dynamics under various climate change scenarios (Caissie et al. (2014)). Deterministic models of water temperature are based on physical principles and are forced by meteorological variables, but are limited by the amount of data required for calibration and by the availability of appropriate models. Stochastic models (Benyahya et al. (2007); Caissie et al. (1998, 2001); Cluis (1972)) are more flexible but may be expensive to evaluate for a given parameter setting. Here we focus on a simple stochastic model for the temporal evolution of river water temperature at a fixed spatial location.

Stochastic simulators of water and air temperature are comprised of an annual trend component and a short-term fluctuation component, or residual. The simula-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

tor requires nearby air temperature data to capture the short-term fluctuations of observed water temperatures. The annual trend is separated from the short term fluctuation by fitting a simple sinusoid to capture annual seasonal variability. Many model formulations have been proposed to capture the residual component, such as Markov models and autoregressive processes (Caissie et al. (1998, 2001)). The model of (Caissie et al., 1998) has

$$T_w(t) = T_a(t) + R_w(t), \tag{2.1}$$

where the annual seasonal component is

$$T_a(t) = a_1 + a_2 \sin\left(\frac{2\pi}{365}(t - t_0)\right),$$

with t as the time index. A simple formulation for the short-term component is related to air temperature residuals as

$$R_w(t) = KR_a(t) + \epsilon,$$

for $\epsilon \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

The calibration parameters $\boldsymbol{\theta} = (a_1, a_2, t_0, K, \sigma)^T$ are the level, a_1, and scaling, a_2, of the annual trend component, the offset term t_0, and the thermal transfer coefficient, K, representing heat transfer from the ambient air into the river water, and σ describes the spread of remaining small-scale variability.

Our observations are temperatures of Alum Creek in Africa, OH (U.S. Geological Survey (2015)) from July 18, 2012 through October 12, 2014. Meterological data is

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

also available (The University of Dayton (2015)), giving the required daily average air temperature data to generate realizations from the stochastic water temperature simulator.

The goal is to estimate the settings of these calibration parameters and predict the state (temperature series) and any model discrepancy between the simulator and observations. We explore calibrating this model to the Alum Creek dataset in Section 4.

2.2 JAK-STAT Model of Intracellular Signaling Pathway

Gene transcription is a complex mechanism that is critical for many biological processes and the understanding of gene transcription in cells is important. Here we describe the JAK-STAT system, a transcription network that has been extensively studied in the literature (Pellegrini and Dusanter-Fourt (1997); Swameye et al. (2003); Timmer et al. (2004); Raue et al. (2009); Horbelt et al. (2002)). Cellular gene transcription begins with a stimulus that is external to the cell. In the JAK-STAT system, the stimulus is the binding of a hormone called Erythropoietin (EpoR_A) to specialized receptors located on the surface of the cell. In response, molecules called transcription factors (Janus kinases), located within the cytoplasm, begin a series of biochemical reactions (phosphorylation) which cycle through an unknown number of reaction states as they move towards the cell nucleus. Once in the nucleus, the transcription factors (now called STATs) begin the process of gene transcription.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

Once completed, the reversible chemical reactions return the chemical species to its original reaction state, allowing the process to begin again. Current understanding of this biochemical reaction includes four reaction states and the possibility of other unknown states proxied by a time delay. The concentrations $x_1(s,\boldsymbol{\theta}),\ldots,x_4(s,\boldsymbol{\theta})$ at time s of the states depend on unknown parameters $\boldsymbol{\theta}\in\mathbb{R}^6$ defined implicitly via the delay differential equation,

$$\begin{cases} \frac{d}{ds}x_1(s,\boldsymbol{\theta}) &= -\ \theta_1\, x_1(s,\boldsymbol{\theta})\ \mathrm{Epo}R_A(s,\boldsymbol{\theta}) + \ 2\,\theta_4\, x_4(s-\theta_5), & s\in[0,60],\\ \frac{d}{ds}x_2(s,\boldsymbol{\theta}) &= \quad \theta_1\, x_1(s,\boldsymbol{\theta})\ \mathrm{Epo}R_A(s,\boldsymbol{\theta}) - \theta_2\, {x_2}^2(s,\boldsymbol{\theta}), & s\in[0,60],\\ \frac{d}{ds}x_3(s,\boldsymbol{\theta}) &= -\ \theta_3\, x_3(s,\boldsymbol{\theta}) + \ \tfrac{1}{2}\,\theta_2\, {x_2}^2(s;\boldsymbol{\theta}), & s\in[0,60],\\ \frac{d}{ds}x_4(s,\boldsymbol{\theta}) &= \quad \theta_3\, x_3(s,\boldsymbol{\theta}) - \theta_4\, x_4(s-\theta_5,\boldsymbol{\theta}), & s\in[0,60],\\ x_1(s;\boldsymbol{\theta}) &= \quad \theta_6, & s\in[-\theta_5,0],\\ x_i(s;\boldsymbol{\theta}) &= \quad 0,\ i=2,3,4, & s\in[-\theta_5,0], \end{cases} \tag{2.2}$$

where subscripts indicate component states. Measurements are made using a process called immunoblotting (Swameye et al. (2003)), which recovers nonlinear transformations of the explicit states contaminated with additive error,

$$\begin{aligned} y_1(s) &= \kappa_1\,(x_2(s)+2x_3(s)) + \epsilon_1(s),\\ y_2(s) &= \kappa_2\,(x_1(s)+x_2(s)+2x_3(s)) + \epsilon_2(s),\\ y_3(s) &= x_1(s) + \epsilon_3(s),\\ y_4(s) &= x_3(s)\,(x_2(s)+x_3(s))^{-1} + \epsilon_4(s), \end{aligned}$$

where the constant multiplicative discrepancies $\boldsymbol{\kappa}=(\kappa_1,\kappa_2)$ reflect the unknown

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

relative scales in the measurement of y_1 and y_2. The errors, $\epsilon_j(s), 1 \leq j \leq 4$, are modeled as independent Gaussian random variables with zero mean and known variances, $\boldsymbol{\lambda_f}^{-1} = (\lambda_{f,1}^{-1}, \ldots, \lambda_{f,4}^{-1})$. Experimental data was obtained from Swameye et al. (2003) and two artificial observations were proposed in Raue et al. (2009) to overcome the lack of identifiability associated with arbitrary units of concentration.The forcing function EpoR_A is modeled by a GP interpolation of its experimental measurements in Swameye et al. (2003).

An important goal is to recover the unknown model parameters and discrepancies, $\boldsymbol{\theta}$ and $\boldsymbol{\kappa}$, based on the measured data $\boldsymbol{y}$. Not only will the rates $\theta_1, \ldots, \theta_4$ help us to understand the phosphorylation reaction, but the delay parameter θ_5 may give an idea of the number of unmodelled states between the fourth state and the original STAT factor. This, in turn, may help future efforts in model building for the JAK-STAT system. Exact inference requires an explicit representation of the concentration states $x_1, \ldots, x_4$, i.e. the solution of model (2.2). For a system of this complexity a solution is not available in closed form. Numerical techniques for delay differential equations suffer from low precision, which has motivated some researchers to replace the above model with a surrogate ordinary differential equation system which is then solved numerically. Our goal here is to use the model (2.2) while accounting for the uncertainty in its numerical solution using the methods of Chkrebtii et al. (2016), but within a constrained amount of computation time.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

3. Model

We now outline the details of our proposed statistical calibration methodology for stochastic simulators with single or multiple states. Due to the high-dimensional nature of our simulator outputs, we consistently use the following conventions:

- index i refers to the ith output grid setting,
- index j refers to the jth setting of the calibration parameter vector,
- index k refers to the kth state output from our multi-state stochastic simulation model, and,
- index u refers to the uth realization of our multi-state stochastic simulation model.

We assume for simplicity that field observations and simulator outputs are available at the same output grid locations $\mathbf{s} \in \mathcal{S}$ for each of the n_s states. Each grid location $\mathbf{s}_i, i = 1, \ldots, n$, is a $p \times 1$ vector representing the setting of p covariate variables. In our applications the $\mathbf{s}_i$ are usually spatial-temporal locations, but this need not be the case. The simulation model takes as input an $\mathbf{s}_i$ and a calibration parameter setting $\boldsymbol{\theta}_j$ resulting in a single realization of the simulator for the kth state being $x_k(\mathbf{s}_i, \boldsymbol{\theta}_j)$.

Our simulation model data consists of an ensemble of such state realizations produced by the computer code given a fixed setting of the parameters $\boldsymbol{\theta}$. We interpret

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

these realizations as i.i.d. samples from some distribution representing uncertainty in the simulation of the process. The stochastic simulators are treated as black-box random functions in the sense that, given inputs, we merely collect realizations from the simulators without any knowledge of internal workings of the stochastic simulators. Our development assumes the availability of N such realizations of the simulation model $x_k(\mathbf{s}_i, \boldsymbol{\theta}_j)$ for each state k at each setting of $\boldsymbol{\theta}_j$ and spatial-temporal location $\mathbf{s}_i$ where N is the number of iterations we require to perform model calibration using our MCMC algorithm. The uth realization of the stochastic simulator $x_k(\mathbf{s}_i, \boldsymbol{\theta}_j)$ is identified as x_{ukij}, $u = 1, \ldots, N$.

In order to emulate stochastic simulators in an approach that is computationally feasible for at least problems of moderate complexity and/or data size, we are motivated by such dimension-reduction ideas as the empirical orthogonal functions (EOFs) (von Storch and Zwiers (1999)) approach to calibration. In the case of stochastic simulators, data dimensionality is much higher, and it does not seem obvious how one should approach the dimension-reduction problem. Our solution is motivated by a tensor representation of our high-dimensional data, which we describe next.

3.1 Tensor Variate Representation of Stochastic Simulator Outputs

The key sources of uncertainty here are the variability across simulator realizations, the variability across states, the variability across the spatial-temporal grid,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

and the variability across calibration parameter settings. A natural way to represent our high-dimensional data is as the $m \times n_s \times n \times N$ multi-dimensional array $\boldsymbol{\chi}$ (Ohlson et al. (2013)). Thus, we express our data as the 4-way tensor $\boldsymbol{\chi} \in \mathbb{R}^{m \times n_s \times n \times N}$. With this representation, the value at tensor entry u, k, i, j is given by $[\boldsymbol{\chi}]_{u,k,i,j} = x_{ukij}$.

Analyzing high-dimensional data structures from the tensor viewpoint has become popular in computer vision (Vasilescu and Terzopoulos (2003)) and Magnetic Resonance Imaging (MRI) applications (Basser and Pajevic (2003)). A key idea in representing high-dimensional data using tensors is how one can decompose the signal in a manner that offers better interpretability. For instance, a D-way tensor can be decomposed into 1-way tensors (vectors) in a procedure analgous to Principal Components Analysis (PCA) performed on a matrix (Lu et al. (2008)). Another approach is the High-Order Singular Value Decomposition (HOSVD) which decomposes a D-way tensor into 2-way tensors (matrices) in a procedure analagous to the SVD of a matrix (Lathauwer et al. (2000a,b)).

The HOSVD tensor decomposition is the more general approach, and is what we use to motivate our model. The HOSVD (Lathauwer et al. (2000a,b); Kolda and Bader (2009)) decomposes our high-dimensional object into a sum of lower-rank objects,

$$[\boldsymbol{\chi}]_{u,k,i,j} = \sum_{r_1}^{R_1} \sum_{r_2}^{R_2} \sum_{r_3}^{R_3} \sum_{r_4}^{R_4} \boldsymbol{\mathcal{E}}_{r_1,r_2,r_3,r_4} a^{(1)}_{u,r_1} a^{(2)}_{k,r_2} a^{(3)}_{i,r_3} a^{(4)}_{j,r_4}, \tag{3.1}$$

where R_1, R_2, R_3, R_4 denote the ranks of the approximation, $\boldsymbol{\mathcal{E}}$ is the $R_1 \times R_2 \times R_3 \times R_4$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

core tensor (analagous to the diagonal weight, or eigenvalue, matrix in the SVD), and $a^{(1)}_{u,r_1} \in \mathbf{A}^{(1)}$ is an entry in the $N \times R_1$ *factor matrix* $\mathbf{A}^{(1)}$, the analogue of an eigenvector in the SVD (similarly for $n_s \times R_2$ matrix $\mathbf{A}^{(2)}$, $n \times R_3$ matrix $\mathbf{A}^{(3)}$ and $m \times R_4$ matrix $\mathbf{A}^{(4)}$). If the entries of $\boldsymbol{\mathcal{E}}$ are Gaussian, then the resulting tensor $\boldsymbol{\chi}$ can be viewed as a draw from a tensor-variate Gaussian Process (Xu et al. (2012)). As such, representing our data as a tensor is the high-dimensional generalization of the GP approach of Kennedy and O'Hagan (2001) and exploiting dimension-reduction techniques for tensors is the high-dimensional generalization of the EOF approach of Higdon et al. (2008).

Equation (3.1) shows that the HOSVD decomposes our tensor object into separate effects arising from variability across simulator runs, variability across states, variability across the spatial-temporal grid, and variability across the stochastic realizations of the simulator. The HOSVD decomposition captures and decomposes the tensor's variability through an operation called *matricization*. Matricization rearranges any tensor into a matrix, and each D-way tensor has D such matricizations. It turns out (Kolda and Bader (2009)) that the dth matricization can be written as

$$\mathbf{X}_{(d)} = \mathbf{A}^{(d)} \boldsymbol{\mathcal{E}}_{(d)} \left(\mathbf{A}^{(D)} \otimes \cdots \otimes \mathbf{A}^{(d+1)} \otimes \mathbf{A}^{(d-1)} \otimes \cdots \otimes \mathbf{A}^{(1)} \right)^T$$

where $\otimes$ represents Kronecker product and $\boldsymbol{\mathcal{E}}_{(d)}$ is the corresponding matricization of the core tensor. Thus, for a D-way tensor in $\mathbb{R}^{I_1 \times \cdots \times I_D}$, the dth matricization re-arranges a tensor into a matrix with I_d rows, stacking the remaining dimensions

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

of the tensor column-wise. For instance, $\mathbf{X}_{(4)}$ matricizes our tensor into a matrix with m rows and $N \times n_s \times n$ columns. The solution to $\mathbf{A}^{(4)}$ in the HOSVD arises as the R_4 left singular vectors from the SVD of $\mathbf{X}_{(4)}$ (Kolda and Bader (2009)). We interpret these left singular vectors as arising from latent eigenfunctions that describe the variability of the tensor across the m simulator runs, motivating the use of a Gaussian process prior.

We reconstruct a missing entry in our tensor representation of simulator outputs, the trajectory of the simulator at the unknown calibration parameter setting $\boldsymbol{\theta}$, by modeling the appropriate eigenvectors. To be clear, we assume that the other matricizations of our tensor are not relevant to our modeling interests. That is, we assume there is no interest in modeling an unobserved state given the observed states; we assume that our data will be on the same grid as the simulator outputs and therefore there is no interest in modeling a state at an off-grid location; and finally, we assume that we have access to all the MCMC realizations of the stochastic simulator required so that modeling a new realization is also not required. Under these assumptions, working with the particular matricization $\mathbf{X}_{(4)}$ is all that is needed to reconstruct the stochastic simulator at the unknown setting $\boldsymbol{\theta}$.

3.2 Modeling Simulator Realizations

Our proposed emulator within the Bayesian hierarchy is constructed as follows. Let $\boldsymbol{\Phi}_u$, $u = 1, \ldots, N$, represent the uth $(n \cdot n_s) \times m$ matrix of simulator realizations of

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

all model states with columns representing the vectors of simulator outputs obtained at the m settings of calibration parameters,

$$\mathbf{\Phi}_u = \begin{bmatrix} x_{u111} & x_{u112} & \dots & x_{u11m} \\ x_{u121} & x_{u122} & \dots & x_{u12m} \\ \vdots & \vdots & \vdots & \vdots \\ x_{u1n1} & x_{u1n2} & \dots & x_{u1nm} \\ \vdots & \vdots & \vdots & \vdots \\ x_{un_sn1} & x_{un_sn2} & \dots & x_{un_snm} \end{bmatrix}.$$

The transpose of matricization $\mathbf{X}_{(4)}$ corresponds to the $(N \cdot n \cdot n_s) \times m$ matrix

$$\mathbf{X}_{(4)}^T = \begin{bmatrix} \mathbf{\Phi}_1 \\ \vdots \\ \mathbf{\Phi}_N \end{bmatrix}.$$

We work with the transpose only so that the matrix orientation follows the typical convention of placing simulator outputs column-wise for each setting of calibration parameters.

Let $\mathbf{X}_{(4)}^T = \check{\mathbf{U}}\check{\mathbf{D}}\check{\mathbf{V}}^T$ be the singular value decomposition (SVD) of $\mathbf{X}_{(4)}^T$, where $\check{\mathbf{U}}$ is $(N \cdot n \cdot n_s) \times m$, $\check{\mathbf{D}}$ is $m \times m$ and $\check{\mathbf{V}}$ is $m \times m$. The low-rank approximation using n_c EOFs from the SVD is $\mathbf{U} = (\mathbf{U}_1, \dots, \mathbf{U}_N)^T$, where each submatrix $\mathbf{U}_u$ is the $(n \cdot n_s) \times n_c$ matrix $\mathbf{U}_u = \tilde{\mathbf{U}}_u \tilde{\mathbf{D}}^{1/2}$, for $u = 1, \dots, N$, with $\tilde{\mathbf{U}}_u$ the $(n \cdot n_s) \times n_c$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

submatrix of $\check{\mathbf{U}}_u$ and $\tilde{\mathbf{D}}$ the $n_c \times n_c$ upper diagonal submatrix of $\check{\mathbf{D}}$. Similarly, let $\mathbf{V} = \tilde{\mathbf{V}}\tilde{\mathbf{D}}^{1/2}$, where $\tilde{\mathbf{V}}$ is the $m \times n_c$ submatrix of $\check{\mathbf{V}}$.

The statistical emulator for each model output is constructed, using the $n_c < m$ EOFs, as

$$x_{ukij} \approx \sum_{l=1}^{n_c} v_l(\boldsymbol{\theta}_j)\mathbf{U}_{ukil}$$

where $v_l(\boldsymbol{\theta}_j) = [\mathbf{V}]_{jl}$ and where the number of bases to use in the approximation, n_c, can be determined by cross-validation as outlined in the Supplementary Materials. This formulation captures some important properties of the chosen EOFs that facilitate the statistical model. Primarily, at different realizations of the simulation model the variation amongst these realizations across states and spatial-temporal locations is completely captured in the left eigenvectors, $\mathbf{U}_u$, while the weights, $v_l(\boldsymbol{\theta}_j)$, do not vary across realizations. This reflects the fact that $\boldsymbol{\theta}_j$ is a fixed, known quantity when the simulator is run at the setting $\boldsymbol{\theta}_j$. Subsequently, it is sensible that the latent weight $v_l(\boldsymbol{\theta}_j)$ should also be treated as fixed conditional on the parameter setting.

For the n_c latent weight spaces, we treat the fixed, known $v_l(\boldsymbol{\theta}_j)$'s and the corresponding unobserved weight $v_l(\boldsymbol{\theta})$ for the unobserved state(s) as realizations of a Gaussian process indexed by the calibration parameter settings,

$$v_l(\boldsymbol{\theta}_1), \ldots, v_l(\boldsymbol{\theta}_m), v_l(\boldsymbol{\theta}) | \lambda_{v_l}, \boldsymbol{\rho}_l, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m, \boldsymbol{\theta} \sim N(0, \lambda_{v_l}^{-1}\mathbf{R}_{v_l}), \tag{3.2}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

where the correlation matrix $\mathbf{R}_{v_l}$ has

$$[\mathbf{R}_{v_l}]_{j,j\prime} = \prod_{t=1}^{q} \rho_{l,t}^{(\theta_{t,j}-\theta_{t,j\prime})^2},$$

for correlation parameters $\boldsymbol{\rho}_l = (\rho_{l,1}, \ldots, \rho_{l,q}) \in (0,1)^q$.

3.3 Modeling Observations

The field observations are modeled as in (1.1). Given the vector of unobserved weights $\mathbf{v}(\boldsymbol{\theta}) = (v_1(\boldsymbol{\theta}), \ldots, v_{n_c}(\boldsymbol{\theta}))^\top$, the additve discrepancies for each state, $\boldsymbol{\delta} = (\boldsymbol{\delta}_1^\top, \ldots, \boldsymbol{\delta}_{n_s}^\top)^\top$, and the multiplicative discrepancies for each state, $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_{n_s})^\top$, the likelihood for each state is

$$\mathbf{y}_k | \mathbf{U}_{uk}, \mathbf{v}(\boldsymbol{\theta}), \boldsymbol{\delta}_k, \kappa_k \sim N(\kappa_k \mathbf{U}_{uk}\mathbf{v}(\boldsymbol{\theta}) + \boldsymbol{\delta}_k, \lambda_{f,k}^{-1}\mathbf{I}_n), \quad k = 1, \ldots, n_s,$$

where $\lambda_{f,k}^{-1}$ corresponds to measurement error variance of the observational data for state k.

3.4 Prior on Discrepancies

Statistical calibration typically accounts for model discrepancy through additive and multiplicative misspecification of the simulator (Kennedy and O'Hagan (2001); Brynjarsdóttir and O'Hagan (2014)), although more general forms have been investigated (Kleiber et al. (2014)). For the additive discrepancy, $\boldsymbol{\delta}$ is modeled using independent GPs for each state variable,

$$\boldsymbol{\delta}_k \sim N\left(\boldsymbol{\mu}_{\delta_k}, \lambda_{\delta_k}^{-1}\mathbf{R}_{\delta_k}\right),$$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

for $k = 1, \ldots, n_s$, where $\lambda_{\delta_k} \sim \text{Gamma}(\alpha_{\delta_k}, \beta_{\delta_k})$, and

$$[\mathbf{R}_{\delta_k}]_{jj'} = \prod_{t=1}^{p} \psi_{k,t}^{(s_{t,j} - s_{t,j'})^2},$$

for $j, j' = 1, \ldots, n$. For the multiplicative discrepancies, we also use independent normal conjugate priors for each state, with $\kappa_k \sim N(\mu_{\kappa_k}, \lambda_{\kappa_k}^{-1})$ for $k = 1, \ldots, n_s$.

Additive discrepancy priors have been discussed in the literature to a reasonable extent (Kennedy and O'Hagan (2001); Higdon et al. (2008); Vernon et al. (2010); Brynjarsdóttir and O'Hagan (2014)) while multiplicative discrepancies are less common. Theoretical aspects of calibration in the presence of model discrepancy has also been explored (Tuo and Wu (2015, 2016)).

3.5 Prior on calibration parameters, θ_t

Assuming calibration parameters have been rescaled to $[0, 1]$, uninformative independent uniform priors are placed on each parameter, $\theta_t \sim \text{Unif}(0, 1)$. When a priori knowledge of the parameters is available, these priors can be adjusted accordingly, as we do in the JAK-STAT example in Section 5.

3.6 Other Prior Distributions

In addition to the main model components – the emulator of Section 3.3, the likelihood of Section 3.4, and the discrepancy priors of Section 3.5 – we need to specify the prior distributions for the remaining unknowns. Generally, specification of these priors is simpler as the model is less sensitive to these parameters unless specified otherwise. We summarize these priors in the Supplementary Materials.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

With all the priors specified as described, the posterior distribution

$$[\{\theta_t\}_{t=1}^{q}, \{v_l(\boldsymbol{\theta})\}_{l=1}^{n_c}, \boldsymbol{\delta}, \boldsymbol{\kappa}, \{\lambda_{v_l}\}_{l=1}^{n_c}, \{\boldsymbol{\rho}\}_{l=1}^{n_c}, \{\lambda_{f,k}\}_{k=1}^{n_s}, \{\lambda_{\delta_k}\}_{k=1}^{n_s}\{\boldsymbol{\psi}_k\}_{k=1}^{n_s}|\mathbf{y}, \mathbf{U}_u, \mathbf{V}] \tag{3.3}$$

$$\propto [\mathbf{y}|\mathbf{U}_u, \mathbf{v}(\boldsymbol{\theta}), \boldsymbol{\delta}, \boldsymbol{\kappa}] \prod_{l=1}^{n_c} ([v_l(\boldsymbol{\theta})|\mathbf{v}_l, \lambda_{v_l}, \boldsymbol{\rho}_l, \boldsymbol{\theta}]\,[\mathbf{v}_l|\lambda_{v_l}, \boldsymbol{\rho}_l, \boldsymbol{\theta}]) \prod_{k=1}^{n_s} [\boldsymbol{\delta}_k|\mu_{\delta_k}, \lambda_{\delta_k}, \boldsymbol{\psi}_k]$$
$$\times \prod_{l=1}^{n_c} ([\lambda_{v_l}]\,[\boldsymbol{\rho}_l]) \prod_{k=1}^{n_s} [\lambda_{f,k}] \prod_{k=1}^{n_s} ([\lambda_{\delta_k}]\,[\boldsymbol{\psi}_k]) \prod_{t=1}^{q} [\theta_t]$$

is sampled via an MCMC algorithm as outlined in the Supplementary Materials.

3.7 Accounting for Simulator Uncertainty

In our framework, simulator uncertainty is propagated through to the statistical calibration by directly sampling from the simulator at the fixed parameter settings $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m$. At each parameter setting $\boldsymbol{\theta}_j$, the corresponding realization is represented by the column vector $\boldsymbol{\Phi}_{uj}$ for $u = 1, \ldots, N$, which we think of as samples from some distribution conditional on the calibration parameter setting $\boldsymbol{\theta}_j$. Each of the N steps in the MCMC algorithm then requires a sample (or, in practice, an approximate sample when $n_c < m$) from this distribution using the basis representation obtained using the $\mathbf{U}_u$'s, $u = 1, \ldots, N$ and $v_l(\boldsymbol{\theta}_j)$'s, $l = 1, \ldots, n_c$, $j = 1, \ldots, m$, from which the uncertainty of the stochastic simulator is propagated through to the statistical calibration.

In EOF-based calibration methods, all inference is conditional on the recovered eigenvectors that are discretely observed over the continuous domain $\mathcal{S}$, the trunca-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

tion of the spectrum of EOF's to a small number of such eigenvectors, and to the particular set of discrete spatial-temporal locations used in observing an otherwise continuous field over $\mathcal{S}$. Such Bayesian models are approximate in that sense, and the approximation improves if the number of spatial-temporal locations goes to infinity (so-called infill asymptotics) and the number of eigenvectors retained goes to infinity. Analogously, inference for the approach outlined is conditional on the additional discrete sampling of N simulator realizations at each of $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m$. These samples discretely approximate the continuous sample space of the conditional distribution of simulator model outputs, and this approximation can be improved by increasing N.

4. Calibrating the Stochastic Water Temperature Model

The water temperature model (2.1) is used to demonstrate the proposed methodology. The model is formed from a seasonal effect component, $T_a(t)$, and a short-term fluctuation component, $R_w(t)$. Many model forms have been proposed to capture the short-term fluctuation component, and the model we use is the simplest, suggesting that some small-scale discrepancy may be present in our calibration.

Plausible ranges for the calibration parameters were chosen by performing an exploratory data analysis, and uniform priors were assigned to each parameter as summarized in Table 1. A set of $m = 30$ calibration parameter settings were chosen using a space-fillling design (Johnson et al. (1990)) and N realizations of the simulator

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

were sampled at each of these settings. Arranging these N realizations column-wise for each setting of the calibration parameters results in our simulator output matrix, $\mathbf{\Phi}_u,\ u = 1, \ldots, N$. Prior distributions for the remaining parameters were chosen according to the default approach described in Sections 3.4 and 3.6. Of particular importance are setting the priors for the discrepancies and λ_f. The water temperature modeling literature suggests that the more complex models for $R_w(t)$ are additive in nature. We assume only additive discrepancies and fix $\kappa = 1$. As the additive discrepancy is expected to account for non-smooth small-scale behaviour, we center the prior mean at $\boldsymbol{\mu}_\delta = \mathbf{0}$ and use the exponential correlation model (Cressie (1993)) for the discrepancy correlation matrix $\mathbf{R}_\delta(\psi)$.

θ	Symbol	Description	Prior
1	a_1	Overall temperature level	Unif(10,20)
2	a_2	Seasonal component scale	Unif(10,20)
3	K	Thermal diffusivity	Unif(0,1)
4	t_0	Seasonal component offset	Unif(50,80)
5	σ	Short-term fluctuation deviation	Unif(0,1)

Table 1: Prior distributions on the calibration parameters for the stochastic water temperature model.

The prior on ψ was chosen to emphasize short-range correlation, taking $\psi \sim$

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

Beta($\alpha_\psi = 1, \beta_\psi = 100$). The scale of the discrepancy was selected to match the 95th percentile (i.e. $\pm$ 2 s.d.) of the range of observed residuals between the observations $\mathbf{y}$ and the first $m = 30$ simulator realizations, $\mathbf{\Phi}_1$. Empirically, the variance of this residual was around 100. Choosing a shape parameter of $\alpha_\delta = 10$, we matched the inverse of the prior mean of λ_δ by re-arranging $\left(\frac{\alpha_\delta}{\beta_\delta}\right)^{-1} = 100$, leading to the prior distribution Gamma($\alpha_\delta = 10, \beta_\delta = 1000$).

The prior on the observational error, λ_f, was selected to match a small percentage, say 10%, of the residual variance calculated above. From this estimate, we arrived at the prior distribution $\lambda_f \sim$ Gamma($\alpha_f = 10, \beta_f = 100$).

The number of components, n_c, retained in the bases expansion was investigated using the leave-one-out cross-validation approach described in the Supplementary Materials. The cross-validated MSPE for predicting the held-out mean simulator and the mean squared error (MSE) of the posterior mean calibration parameter estimates (scaled to unit interval) are summarized in Table 2. The results of this cross-validation study suggest $n_c = 4$ bases is a good compromise between accuracy and computational cost.

The results of calibrating the water temperature simulator to the Africa, OH dataset are shown in Figures 1 and 2. The emulator (black lines) in Figure 1 fits the data well, demonstrating good coverage of the observed data (red dots), yet there is clear evidence of a small-scale discrepancy (blue lines). The presence of this

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

	$n_c = 2$	$n_c = 3$	$n_c = 4$	$n_c = 5$	$n_c = 6$	$n_c = 8$	$n_c = 10$
MSPE	2.74	2.71	0.044	0.018	0.067	0.048	0.051
MSE($\boldsymbol{\theta}$)	0.118	0.097	0.047	0.065	0.076	0.072	0.058

Table 2: Effect of varying the number of bases, n_c, used in the model on the cross-validated MSPE of the mean held-out state and MSE of the estimated calibration parameters.

discrepancy, which appears discontinuous and autocorrelated, is in agreement with the assumptions found in more advanced models of $R_w(t)$ in the literature, such as AR(1) and AR(2) models (Caissie et al. (1998)).

The MCMC algorithm for the proposed calibration model was iterated for $N = 50,000$ steps, with the first $25,000$ being discarded as burn-in. The posterior densities for the calibration parameters shown in Figure 2 indicate that most parameters are well determined despite the stochasticity of the simulator. However, the thermal diffusivity coefficient, K, is less well determined than the annual model component parameters a_1, a_2, and t_0. This is not suprising given the presence of discrepancy and the underparameterized form of $R_w(t)$ employed. With a more flexible model of the small-scale structure for $R_w(t)$, the diffusivity coefficient might be better resolved.

In comparison, fitting the model using a single realization of the simulator showed notable differences, particularly in the assessment of uncertainties. For instance,

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

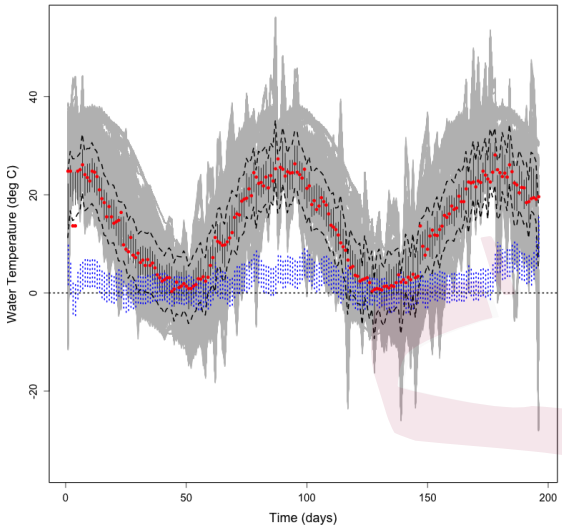

Figure 1: 25,000 posterior samples of the calibrated stochastic simulator and uncertainties. The grey lines represent the prior realizations of the stochastic simulator, while the vertical black lines represent the 95% credible intervals for the calibrated discrepancy-corrected simulator and the dashed lines represent 95% credible intervals for the observed process. The vertical dotted lines are 95% credible intervals for the additive discrepancy component, which is at the level of zero (horizontal dotted line) but does exhibit small-scale structure as expected.

the standard deviations of the posterior distributions of the calibration parameters as summarized in Table 3 were consistently smaller when accounting for simulator uncertainty as compared to deterministic calibration. This suggests that accounting

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

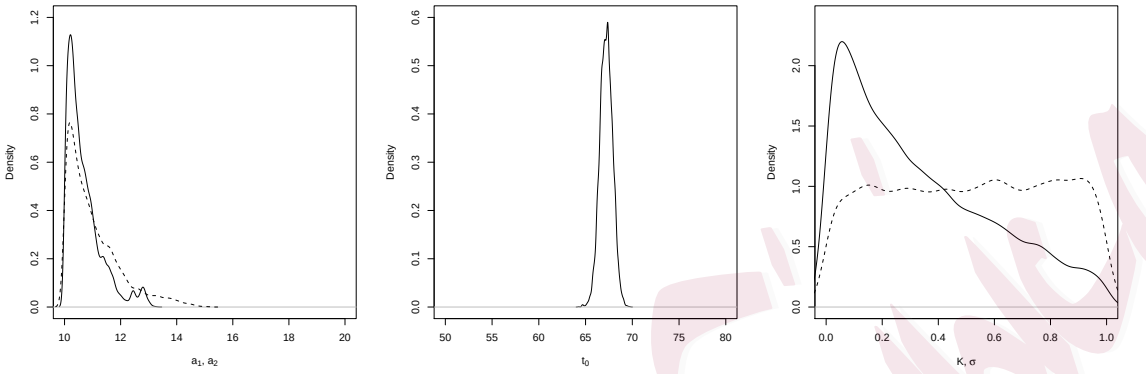

Figure 2: Kernel density estimates for the calibration parameters of the water temperature stochastic simulator based on 25,000 posterior samples. Calibration paramerers a_1 (solid) and a_2 (dashed) shown in the left pane, t_0 shown in center pane, K (solid) and σ (dashed) shown in right pane.

for simulator uncertainty can actually remove uncertainty that might otherwise be mapped to parameter estimates when performing deterministic calibration.

The correlations of the estimated parameters shown in Table 4 are not captured when assuming the simulator is deterministic. This can be important information for interpreting the simulator or performing variable selection. Moreover, this suggests that the joint distribution of the calibration parameters is better determined when the stochasticity of the simulator is accounted for as compared to a deterministic analysis.

While the predictions of both models are good (as one would expect since both

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

	a_1	a_2	K	t_0	σ
Stochastic Calibration	0.045	0.062	0.218	0.025	0.291
Deterministic Calibration	0.056	0.109	0.263	0.032	0.286

Table 3: Sample standard deviations of posterior calibration parameter realizations using stochastic versus deterministic calibration mdels.

	a_1	a_2	K	t_0	σ
a_1	1.00	-0.72 (-0.13)	0.17 (-0.13)	0.23 (0.02)	-0.02 (-0.01)
a_2	-0.72 (-0.13)	1.00	-0.26 (-0.09)	-0.32 (-0.13)	-0.01 (0.01)
K	0.17 (-0.13)	-0.26 (-0.09)	1.00	0.15 (0.28)	0.01 (0.00)
t_0	0.23 (0.02)	-0.32 (-0.13)	0.15 (0.28)	1.00	-0.02 (0.08)
σ	-0.02 (-0.01)	-0.01 (0.01)	0.01 (0.00)	-0.02 (0.08)	1.00

Table 4: Pearson correlations of parameters estimated using the stochastic versus deterministic (in brackets) calibration models.

models include discrepancies), the second order properties again show some differences. For instance, the standard deviations of the posterior distributions for the predicted process are similar for both models, but the standard deviation for the discrepancy when accounting for simulator uncertainty (0.776) was about 10% smaller than when performing deterministic calibration (0.862). Similarly, the standard deviation of the posterior emulated state, when accounting for simulator uncertainty

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

(0.607), was about 17% smaller than when performing deterministic calibration (0.732). Taken together, these results demonstrate that accounting for uncertainty in stochastic simulators can lead to more efficient uncertainty quantification in the resulting calibration.

5. Calibrating the JAK-STAT Model

To show how we enable calibration of probabilistic differential equation solvers that capture state discretization uncertainty as part of the probabilistic solution, we study the JAK-STAT system described in Section 2. Because (2.2) has no closed form solution, Chkrebtii et al. (2016) performed exact inference by directly modeling uncertainty associated with discretization of the states within the inverse problem. The drawback of this approach is the computational expense incurred. A computer model calibration approach could significantly reduce the computational cost, but must account for the stochastic nature of the probabilistic solver. We overcome this difficulty by making calibration for probabilistic solution simulators feasible.

Our goal is to infer calibration parameters $\boldsymbol{\theta} \in \mathbb{R}^8$ where θ_1 through θ_4 represent reaction rates in model (2.2), θ_5 is the time required for the process of gene transcription to begin and for the reaction states to return to the original state, θ_6 is the initial concentration for the first reaction state, and θ_7 and θ_8 are unknown hyperparameters associated with the probabilistic solver. Prior distributions on the calibration parameters are provided in Table 5. For a given parameter regime, the

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

model of discretization uncertainty of (2.2) produces posterior draws based on an equally-spaced time discretization grid of size 500. The emulator is constructed from a random design with $m = 100$ calibration parameter settings drawn from the prior distributions in Table 5. For this application, we expect that fine scale structure in the state may not be captured by using a small number of parameter settings to construct the emulator, therefore an additive model discrepancy, $\boldsymbol{\delta}$, is introduced on the observation process as described in Section 3.3. It is assigned a Gaussian process prior with stationary squared exponential covariance structure and zero prior mean $\boldsymbol{\mu}_\delta$. The prior model on the precision parameters $\boldsymbol{\lambda}_\delta, \boldsymbol{\lambda}_f$ is described in Sections 3.4 and 3.6.

Symbol	Description	Prior
$\theta_i,\ i = 1, \ldots, 4$	Reaction rates of first for states	χ^2_1
θ_5	Time delay	χ^2_6
θ_6	Initial concentration of the first state	$\mathrm{N}\left(y^{(3)}(0), 40^2\right)$
θ_7	Prior precision of the probabilistic solver	$100 + \text{Log-N}\,(10, 1)$
θ_8	Length-scale of the probabilistic solver	$0.12 + \mathrm{Exp}\,(0.1)$

Table 5: Prior distributions on the calibration parameters for the JAK-STAT system.

Our analysis is based on 20,000 posterior samples. The marginal posteriors over the observation processes are superimposed on the data in Figure 4, and fit

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

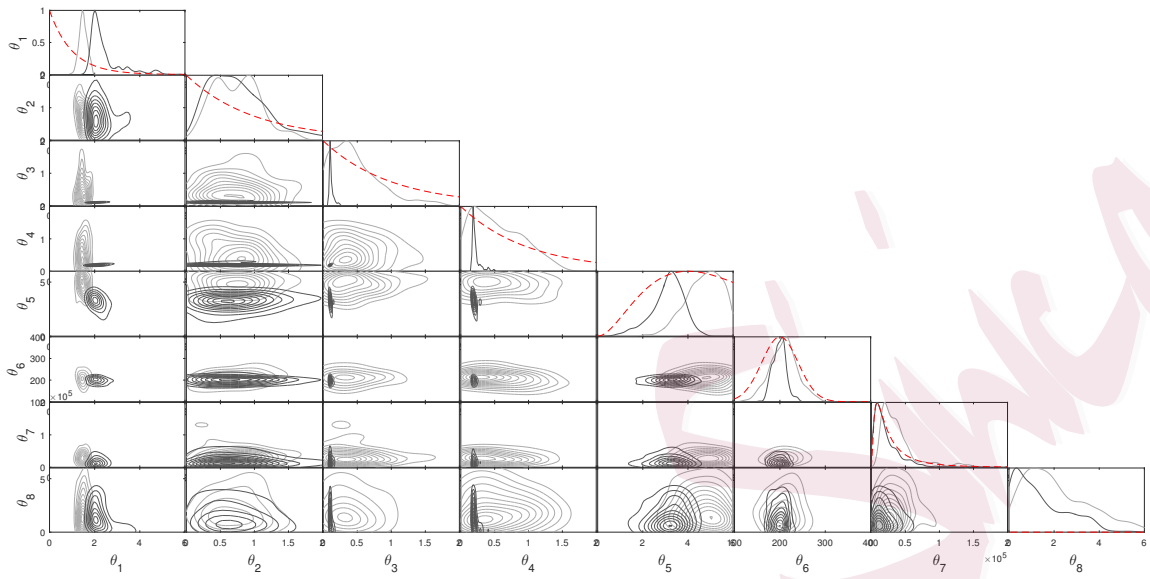

Figure 3: Kernel density estimates of the marginal stochastically calibrated posterior (gray) with $m = 100$ model runs, and exact posterior (black, Chkrebtii et al. (2016)) for the JAK-STAT system. Marginal prior densities are shown as dotted lines.

well overall without fully capturing all the small scale structure, as expected. The discrepancy captures structure that is not contained in the emulated model space, including any misspecification in the original delay differential equation model for the system itself. We find that the discrepancy captures very little structure, and is rather diffuse and essentially stationary. Interestingly, increasing the number, m, of sampled parameter settings from 20 and 50 (shown in the Supplementary Materials) to 100 had little effect on the fit of the model to the data and structure of the discrepancy although with a noticeable decrease in uncertainty. This suggests that

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

parameter regions of very high posterior probability may be small relative to scale of parameter sampling regions, as expected for such highly nonlinear problems.

Increasing the number of model runs does, however, push the posterior density of several rate parameters further from the prior (posterior density plots for $m = 20, 50$ computer model runs are provided in the Supplementary Materials). Estimated marginal posterior densities for the calibration parameters are shown in Figure 3. Based on observed differences from the marginal priors, we conclude that the data is informative for all the parameters except for the initial concentration, θ_6, of the first state, which depends on the scaling of the concentration units and is not identifiable given the experimental data (e.g. Raue et al. (2009)). Further, the marginal posterior distributions of the calibrated model are more diffuse than their exact counterparts in Chkrebtii et al. (2016), due to the contribution of uncertainty from emulating the exact model based on a finite number, m, of model evaluations. Despite this, the posterior modes align well with their exact analogues while computational gain is dramatic. Indeed, performing the calibration using the proposed method requires about 30 minutes on a modern notebook computer for 20,000 samples of the posterior, while the same number of samples using the full solution method of Chkrebtii et al. (2016) requires over a day.

6. Discussion

In this paper we have presented an approach for calibration stochastic simula-

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

tors in a computationally efficient manner while allowing for the uncertainty in the simulator outupts to be propagated through to the calibration parameter estimates as well as the state and discrepancy predictions. Our method also allows for multiple states to be calibrated simultanenously within the same framework. The proposed model can thus be viewed as a higher-dimensional generalization of the deterministic, single-state EOF-based approach to calibration first described in Higdon et al. (2008).

Applying the methodology to our two motivating examples suggests that accounting for the non-determinism in some simulators can be important. In the water temperature example, a simple stochastic simulator of water temperature captures seasonal variability through a functional form and small-scale structure through a thermal diffusivity model that connects ambient air temperature data to the water temperature. The proposed method provided plausible estimates of the model parameters while capturing expected discrepancy in the model for diffusivity due to the underparameterized form of the small-scale structure used. The discrepancy found is in agreement with more complex models of thermal diffusivity found in the river water temperature modeling literature. In comparison, deterministic calibration underestimated pairwise correlations of calibration parameters and had wider uncertainties for most estimated quantities. This suggests that accounting for the stochasticity of the simulator more accurately captures the full joint distribution of

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

parameters.

In the second example, the proposed methodology enables emulation of Bayesian probability models of discretization uncertainty in the solution of differential equations. We have demonstrated its feasibility and computational efficiency on the complex JAK-STAT gene trancription network. The resulting posterior parameter distributions as well as state and discrepancy estimates are largely in close agreement with the exact method found in the literature. Yet, the computational cost is vastly reduced. This result is a promising step forward in extending the scope of discretization uncertainty modeling to possibly large-scale systems, such as those used in oceanography and atmospheric sciences, where small perturbations in the state, such as those due to discretization as well as model discrepancy, can have a substantial impact.

One limitation of the model described is the possibility of additional inputs, $\mathbf{z}$, which can be controlled both in the simulation model and in the real-world. A common example of this situation is that of settings of temperature and pressure in engineering applications. When outputs and observations are available on the same spatial-temporal grid for each setting of the joint input parameters $(\boldsymbol{\theta}, \mathbf{z})$, our approach can easily accomodate this by including the additional variables $\mathbf{z}_1, \ldots, \mathbf{z}_m$ and $\mathbf{z}$ in the GP emulation model in (3.2), recognizing that the setting of these parameters for the field data, $\mathbf{z}$, is fixed and known. More generally, our model cannot

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

accomodate spatial-temporal inputs that are not crossed with the input settings $(\boldsymbol{\theta}, \mathbf{z})$. Here, the conceptual framework of tensors and matricization introduced in Section 3 suggests the possibility of modeling more than one matricization in order to emulate the desired outputs in such a scenario. Another possible extension would be the case of multiple simulation models which has been addressed by Bayesian Model Averaging techniques (Raftery et al. (2005); Hoeting et al. (1999)). Combining all three sources of uncertainty - multiple simulators, simulator emulation and stochastic simulators - would be a challenging but potentially very interesting endeavor.

We have developed a Bayesian model for calibrating complex multi-state non-deterministic simulators. Our treatment of simulator stochasticity is more honest than assuming a simple i.i.d. error (nugget) model, yet the approach only relies on samples of the simulator being available rather than knowledge of its full distribution in closed-form, which is typically unavailable. The method is implemented in `R` (R Core Team (2012)) and will shortly be available as package `cmce` on `CRAN`.

Supplementary Materials

The online supplement contains additional details of the MCMC algorithm, calibrating prior distributions, and additional figures for the JAK-STAT example.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

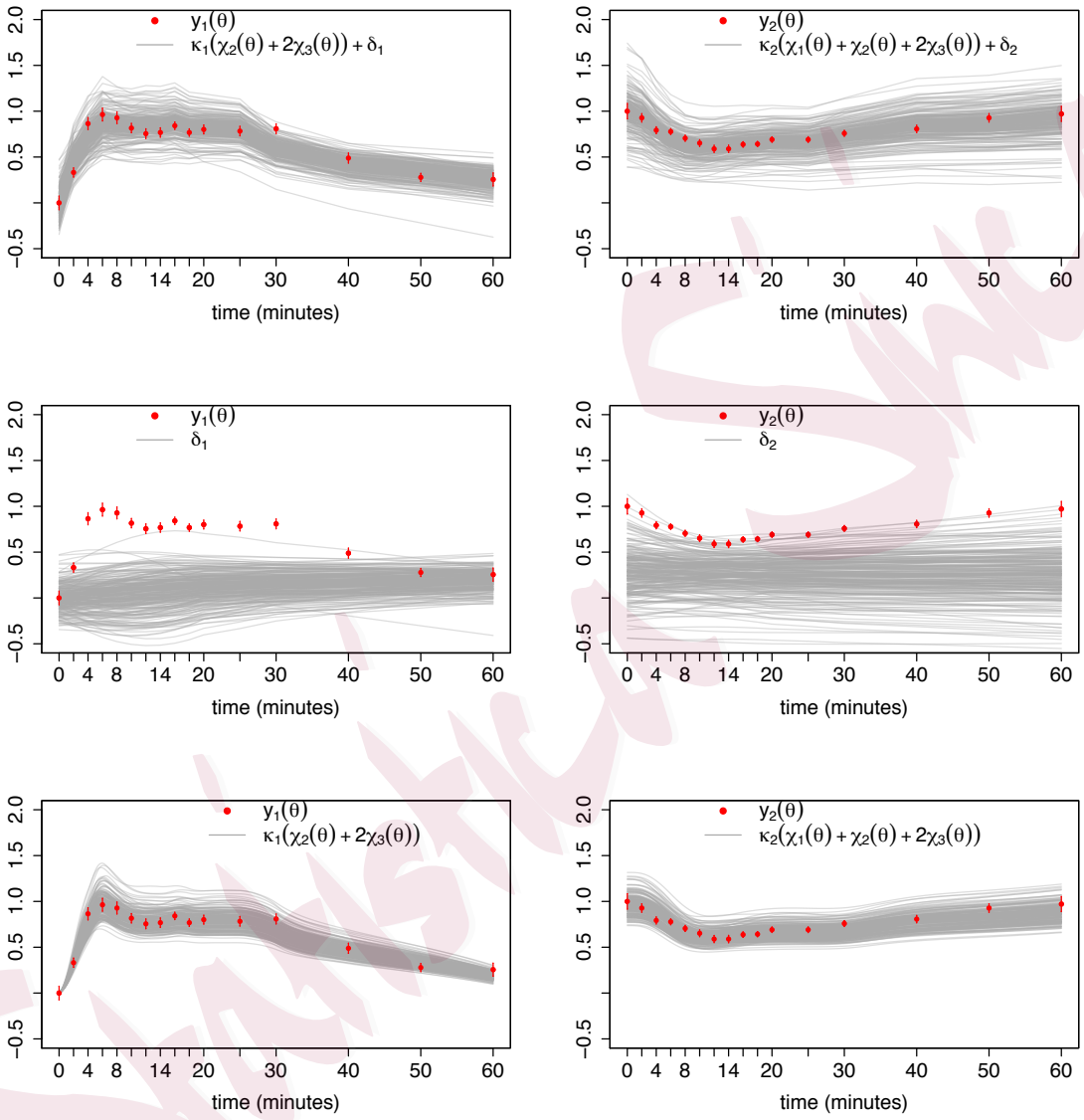

Figure 4: 200 samples from the marginal calibrated posterior with $m = 100$ model runs (top row), discrepancies δ_1 and δ_2 (middle row), and exact posterior for comparison (bottom row, Chkrebtii et al. (2016)) over the first two observation processes of the JAK-STAT system, for which experimental data is available. Experimental data locations are shown as red circles with error bars representing twice the experimental error standard deviation.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

References

Arridge, S. R., J. P. Kaipio, V. Kolehmainen, M. Schweiger, E. Somersalo, T. Tarvainen, and M. Vauhkonen (2006). Approximation errors and model reduction with an application in optical diffusion tomography. *Inverse Problems 22*, 175–195.

Auchincloss, A. H. and A. V. D. Roux (2008). A new tool for epidemiology: the usefulness of dynamic-agent models in understanding place effects on health. *American journal of epidemiology 168*, 1–8.

Basser, P. J. and S. Pajevic (2003). A normal distribution for tensor-valued random variables: Applications to diffusion tensor MRI. *IEEE Transactions on Medical Imaging 22*, 785–794.

Benyahya, L., D. Caissie, A. St-Hilaire, T. Ouarda, and B. Bobe'ee (2007). A review of statistical water temperature models. *Canadian Water Resources Journal 32*, 179–192.

Brynjarsdóttir, J. and A. O'Hagan (2014). Learning about physical parameters: The importance of model discrepancy. *Inverse Problems 30*, 114007.

Caissie, D., N. El-Jabi, and A. St-Hilaire (1998). Stochastic modelling of water temperatures in a small stream using air to water relations. *Canadian Journal of Civil Engineering 25*, 250–260.

Caissie, D., N. El-Jabi, and N. Turkkan (2014). Stream water temperature modeling under climate change scenarios B1 & A2. Technical report, Canadian Tech Report of Fisheries and Aquatic Sciences, Moncton, New Brunswick.

Caissie, D., N. El-Jabib, and M. G. Satish (2001). Modelling of maximum daily water temperatures in a small stream using air temperatures. *Journal of Hydrology 251*, 14–28.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

REFERENCES

Chkrebtii, O., D. A. Campbell, B. Calderhead, and M. Girolami (2016). Bayesian solution uncertainty quantification for differential equations. *Bayesian Analysis 11*, 1239–1267.

Cluis, D. (1972). Relationship between stream water temperature and ambient air temperature - a simple autoregressive model for mean daily stream water temperature fluctuations. *Nordic Hydrology 3*, 65–71.

Cressie, N. A. C. (1993). *Statistics for Spatial Data, rev. edn.* New York: John Wiley & Sons Inc.

Gilbert, G. N. (2008). *Agent-based models.* Sage Publications, Inc.

Goldstein, M. and J. Rougier (2006). Bayes linear calibrated prediction for complex systems. *Journal of the American Statistical Association 101*, 1132–1143.

Higdon, D., J. Gattiker, B. Williams, and M. Rightley (2008). Computer model calibration using high-dimensional output. *Journal of the American Statistical Association 103*, 570–583.

Higdon, D., M. C. Kennedy, J. Cavendish, J. Cafeo, and R. D. Ryne (2004). Combining field data and computer simulations for calibration and prediction. *Siam Journal on Scientific Computing 26*, 448–466.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: a tutorial. *Statistical Science 14*, 382–401.

Horbelt, W., J. Timmer, and H. U. Voss (2002). Parameter estimation in nonlinear delayed feedback systems from noisy data. *Physics Letters A 299*, 512–521.

Johnson, M. E., L. M. Moore, and D. Ylvisaker (1990). Minimax and maximin distance designs. *Journal*

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

of Statistical Planning and Inference 26, 131–148.

Joseph, V. R. and S. N. Melkote (2009). Statistical adjustments to engineering models. *Journal of Quality Technology 41*, 362–375.

Kennedy, M. and A. O'Hagan (2001). Bayesian calibration of computer models (with discussion). *Journal of the Royal Statistical Society: Series B 68*, 425–464.

Kim, J. H., T. Abel, O. Agertz, G. L. Bryan, D. Ceverino, C. Christensen, and J. Guedes (2013). The agora high-resolution galaxy simulations comparison project. *The Astrophysical Journal Supplement Series 210*, 14.

Kleiber, W., S. R. Sain, and M. J. Wiltberger (2014). Model calibration via deformation. *SIAM/ASA Journal on Uncertainty Quantification 2*, 545–563.

Kolda, T. G. and B. W. Bader (2009). Tensor decompositions and applications. *SIAM Review 51*, 455–500.

Lathauwer, L. D., B. D. Moor, and J. Vandewalle (2000a). A multilinear singular value decomposition. *SIAM Journal of Matrix Analysis and Applications 21*, 1253–1278.

Lathauwer, L. D., B. D. Moor, and J. Vandewalle (2000b). On the best rank-1 and rank-(R1, R2, ..., Rn) approximation of higher-order tensors. *SIAM Journal of Matrix Analysis and Applications 21*, 1324–1342.

Lu, H., K. N. Plataniotis, and A. N. Venetsanopoulos (2008). MPCA: Multilinear principal component analysis of tensor objects. *IEEE Transactions on Neural Networks 19*, 18–39.

Ohlson, M., M. R. Ahmad, and D. Von Rosen (2013). The multilinear normal distribution: Introduction

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

and some basic properties. *Journal of Multivariate Analysis 113*, 37–47.

Palmer, R. G., W. B. Arthur, J. H. Holland, B. LeBaron, and P. Tayler (1994). Artificial economic life: a simple model of a stockmarket. *Physica D: Nonlinear Phenomena 75*, 264–274.

Pellegrini, S. and I. Dusanter-Fourt (1997). The structure, regulation and function of the janus kinases (JAKs) and the signal transducers and activators of transcription (STATs). *European Journal of Biochemistry 248*, 615–633.

R Core Team (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005). Using bayesian model averaging to calibrate forecast ensembles. *Monthly Weather Review 133*, 1155–1174.

Raue, A., C. Kreutz, T. Maiwald, J. Bachmann, M. Schilling, U. Klingmüller, and J. Timmer (2009). Structural and practical identifiability analysis of partially observed dynamical models by exploiting the profile likelihood. *Bioinformatics 25*, 1923–1929.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn (1989). Design and analysis of computer experiments (with discussion). *Statistical Science 4*, 409–423.

Swameye, I., T. G. Muller, J. Timmer, O. Sandra, and U. Klingmuller (2003). Identification of nucleocytoplasmic cycling as a remote sensor in cellular signaling by databased modeling. *Proceedings of the National Academy of Sciences 100*, 1028–1033.

Tesfatsion, L. (2002). Agent-based computational economics: Growing economies from the bottom up.

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

Artificial life 8, 55–82.

The University of Dayton (2015). Environmental protection agency average daily temperature archive. `http://academic.udayton.edu/kissock/http/Weather/citylistUS.htm`. Accesssed: 2015-07-20.

Timmer, J., T. G. Muller, I. Swameye, O. Sandra, and U. Klingmuller (2004). Modeling the nonlinear dynamics of cellular signal transduction. *International Journal of Bifurcation and Chaos 14*, 2069–2079.

Tuo, R. and C. F. J. Wu (2015). Efficient calibration for imperfect computer models. *The Annals of Statistics 43*, 2331–2352.

Tuo, R. and C. F. J. Wu (2016). A theoretical framework for calibration in computer models: parametrization, estimation and convergence properties. *SIAM/ASA Journal on Uncertainty Quantification 4*, 767–795.

U.S. Geological Survey (2015). Usgs water data for the nation. `http://waterdata.usgs.gov/`. Accessed: 2015-07-20.

Vasilescu, M. A. O. and D. Terzopoulos (2003). Multilinear subspace analysis for image ensembles. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR 2003)*, Volume 2, pp. 93–99. IEEE Computer Society.

Vernon, I., M. Goldstein, and R. G. Bower (2010). Galaxy formation: A bayesian uncertainty analysis. *Bayesian Analysis 5*, 619–670.

von Storch, H. and F. W. Zwiers (1999). *Statistical Analysis in Climate*. New York: Cambridge University

Statistica Sinica: Preprint
doi:10.5705/ss.202016.0403

Press.

Xu, Z., F. Fan, and A. Qi (2012). Infinite tucker decomposition: Nonparametric Bayesian models for multiway data analysis. In *Proceedings of the 29th International Conference on Machine Learning*.

Department of Statistics, The Ohio State University

E-mail: (mpratola@stat.osu.edu)

Department of Statistics, The Ohio State University

E-mail: (oksana@stat.osu.edu)