Operations Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

A Bayesian Framework for Quantifying Uncertainty in Stochastic Simulation

Wei Xie, Barry L. Nelson, Russell R. Barton

OPERATIONS RESEARCH
Vol. 62, No. 6, November–December 2014, pp. 1439–1452
ISSN 0030-364X (print) | ISSN 1526-5463 (online)

http://dx.doi.org/10.1287/opre.2014.1316
© 2014 INFORMS

A Bayesian Framework for Quantifying Uncertainty in Stochastic Simulation

Wei Xie
Department of Industrial and Systems Engineering, Rensselaer Polytechnic Institute, Troy, New York 12180, xiew3@rpi.edu

Barry L. Nelson
Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, Illinois 60208, nelsonb@northwestern.edu

Russell R. Barton
Smeal College of Business, Pennsylvania State University, University Park, Pennsylvania 16802, rbarton@psu.edu

When we use simulation to estimate the performance of a stochastic system, the simulation often contains input models that were estimated from real-world data; therefore, there is both simulation and input uncertainty in the performance estimates. In this paper, we provide a method to measure the overall uncertainty while simultaneously reducing the influence of simulation estimation error due to output variability. To reach this goal, a Bayesian framework is introduced. We use a Bayesian posterior for the input-model parameters, conditional on the real-world data, to quantify the input-parameter uncertainty; we propagate this uncertainty to the output mean using a Gaussian process posterior distribution for the simulation response as a function of the input-model parameters, conditional on a set of simulation experiments. We summarize overall uncertainty via a credible interval for the mean. Our framework is fully Bayesian, makes more effective use of the simulation budget than other Bayesian approaches in the stochastic simulation literature, and is supported with both theoretical analysis and an empirical study. We also make clear how to interpret our credible interval and why it is distinctly different from the confidence intervals for input uncertainty obtained in other papers.

Subject classifications: input uncertainty; Bayesian inference; credible interval; simulation output analysis; metamodel; Gaussian process.
Area of review: Simulation.
History: Received October 2013; revisions received March 2014, July 2014; accepted August 2014. Published online in *Articles in Advance* October 21, 2014.

1. Introduction

Stochastic simulation is used to characterize the behavior of complex, dynamic systems that are driven by random input processes. The distributions of these input processes are often estimated from real-world data. Thus, there are at least two sources of uncertainty in simulation-based estimates: input estimation error—due to only having a finite sample of real-world data—and simulation estimation error—due to only expending a finite amount of simulation effort. Of course, the logic of the simulation model itself may also be wrong, but that is not the focus of this paper. See Chapter 5 in Nelson (2013) for a comprehensive description of simulation errors.

There are already robust methods for quantifying the simulation estimation error. A formal quantification of input estimation error, however, is rarely obtained, and no simulation software routinely does it. Since input estimation error can overwhelm simulation error (Barton et al. 2014), ignoring it may lead to unfounded confidence in the assessment of system performance, which could be the basis for critical and expensive decisions. Thus, it is desirable to quantify the overall impact of simulation and input uncertainty on system performance estimates. Although we focus on the system mean response, our methods can be extended to other performance estimates, such as variances and probabilities.

In this paper we address problems with univariate, parametric input models that are mutually independent and with input-model parameters estimated from a finite sample of real-world data, denoted generically by $\mathbf{z}_\mathbf{m}$, where $\mathbf{m}$ is a vector whose elements are the number of real-world observations available for each input process. This implies that the input models are uniquely specified by their parameters, denoted generically by $\boldsymbol{\theta}$. Let $\mu(\boldsymbol{\theta})$ be the true simulation mean response given parameters $\boldsymbol{\theta}$; that is, $\mu(\cdot)$ is an unknown function that maps parameters of the input distributions into the expected value of the simulation output. If $\boldsymbol{\theta}^c$ denotes the unknown true parameters, then the goal of the simulation is to estimate the true mean response $\mu^c \equiv \mu(\boldsymbol{\theta}^c)$. We want to quantify the overall estimation uncertainty about μ^c while simultaneously reducing the uncertainty introduced during the propagation from inputs to outputs.

There are various methods proposed in the literature to quantify the uncertainty due to estimating input-model parameters, which we call *input uncertainty*; see

Barton (2012) for a review. The methods can be divided into frequentist and Bayesian approaches. The frequentist approaches start with a point estimate of the input-model parameters, $\hat{\boldsymbol{\theta}}$, which is a function of real-world data $\mathbf{z_m}$. Since the real-world data are one of many possible random samples, the uncertainty about $\hat{\boldsymbol{\theta}}$ is quantified by its sampling distribution. The input-parameter uncertainty is then propagated to the output mean through direct simulation or a metamodel, either of which introduces additional uncertainty. For any fixed $\boldsymbol{\theta}$, let $\hat{\mu}(\boldsymbol{\theta})$ be a point estimate of the system mean response. One way to summarize the overall estimation uncertainty for μ^c is to invert the sampling distribution of $\hat{\mu}(\hat{\boldsymbol{\theta}})$ and get a $(1-\alpha)100\%$ confidence interval (CI), denoted by $[C_L, C_U]$, such that

$$\Pr\{\mu^c \in [C_L, C_U]\} = 1-\alpha.$$

What difficulties arise when we use the frequentist approaches? First, it may not be possible to obtain the sampling distribution of $\hat{\boldsymbol{\theta}}$. Thus, asymptotic results are often invoked to approximate it; two of these are the normal approximation and the bootstrap. Their validity requires large samples of real-world data. However, "large" is relative and it depends on the input models and the values of the parameters. Thus, the finite-sample performance of these approximations could vary for different stochastic systems. In addition, it is difficult for the frequentist methods to incorporate prior information about the input-model parameters.

A Bayesian approach avoids some of these issues while raising others. Bayesians represent the uncertainty in our belief about $\boldsymbol{\theta}^c$ via a random vector $\boldsymbol{\Theta}$.[1] Before collecting any real-world data, our belief is quantified by its prior distribution $\pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$. After observing the real-world data $\mathbf{z_m}$, our belief is updated using the assumed parametric distribution family of the data and Bayes' rule to yield a posterior distribution denoted by $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m})$. The posterior of $\boldsymbol{\Theta}$ corresponds to the sampling distribution of $\hat{\boldsymbol{\theta}}$ in the frequentist approaches in that both of them characterize the input-parameter uncertainty. However, Bayesians have a fundamentally different perspective on quantifying uncertainty, and answer different questions; correctly and consistently capturing this perspective is one motivation for our work.

The distribution $\pi_{\boldsymbol{\Theta}}$ provides a convenient way to account for prior information about the input-model parameters, if we have any. If not, a noninformative prior can be used. There is no need to rely on a large-sample asymptotic approximation to the sampling distribution. When the real-world sample size is small, then the variance of the posterior distribution will be large. However, evaluation of the posterior distribution can be difficult, so computational approaches, such as Markov Chain Monte Carlo (MCMC), may be needed.

In this paper we take a Bayesian approach to quantify the uncertainty about μ^c. To that end, we let $\boldsymbol{\Theta}$ be a random variable whose distribution represents our knowledge of $\boldsymbol{\theta}^c$. Similarly, we let $M(\cdot)$ be a random function (also called a random field) whose distribution represents our knowledge of $\mu(\cdot)$. The domain of $M(\cdot)$ is the same as that of $\mu(\cdot)$, which is the natural space of feasible values for the input parameters $\boldsymbol{\theta}$ for the input distributions in use. The distribution of $M(\cdot)$ is characterized through the joint distribution of any finite collection $\{M(\boldsymbol{\theta}_1), M(\boldsymbol{\theta}_2), \ldots, M(\boldsymbol{\theta}_p)\}$, which will be Gaussian in our case. See Chapter 1 in Adler (2010) for the existence of the distribution of a random field. To reduce the uncertainty about $\boldsymbol{\Theta}$ and $M(\cdot)$, we employ real-world input data and simulation experiments, respectively, along with Bayes' rule. To represent the overall estimation uncertainty for μ^c, we want to make statements about the composite random variable $U \equiv M(\boldsymbol{\Theta})$.

Bayesian quantification of the uncertainty about $\boldsymbol{\Theta}$ is completely standard. Our interest is in uncertainty about U. If the response function $\mu(\cdot)$ were known, then the impact of input uncertainty on the system mean response could be characterized by an induced posterior distribution for U:

$$F_U(u \mid \mathbf{z_m}, \mu(\cdot)) \equiv \Pr\{\mu(\boldsymbol{\Theta}) \leqslant u \mid \mathbf{z_m}\} \quad \text{with } \boldsymbol{\Theta} \sim p_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m}). \tag{1}$$

From this we could construct a $(1-\alpha)100\%$ credible interval (CrI) for U, denoted by $[Q_L, Q_U]$, which contains $1-\alpha$ of the probability content: $F_U(Q_U \mid \mathbf{z_m}, \mu(\cdot)) - F_U(Q_L \mid \mathbf{z_m}, \mu(\cdot)) = 1-\alpha$. Since there is not a unique CrI meeting this requirement, we use a two-sided, equal-tail probability $(1-\alpha)100\%$ CrI for illustration in this paper. Our approach can also be extended to other criteria, e.g., the highest posterior density CrI. Notice that the CrI depends not only on the data $\mathbf{z_m}$ but also on the prior distribution $\pi_{\boldsymbol{\Theta}}$, which means different analysts with the same data could have different, but completely valid, CrIs. Further, the quality of the CrI is not based on "coverage" but rather on whether it correctly reflects the remaining uncertainty about U after accounting for available information via Bayes' rule.

In reality $\mu(\cdot)$ is unknown, so we have to estimate the mean response; this introduces additional uncertainty. For this reason we refer to $[Q_L, Q_U]$ derived from (1) as the "perfect fidelity" CrI that we could obtain without observing more real-world data because the input uncertainty is propagated to the output mean using the true mean response without introducing any additional error; it provides the standard against which we compare our method and other Bayesian approaches in the simulation literature.

There are two central contributions of the paper: First, we provide a fully Bayesian framework to quantify uncertainty about U along with a method to realize a CrI based on it. Second, we show that our Bayesian framework makes effective use of the computational budget as measured by closeness of our CrI to the perfect fidelity CrI.

Our framework represents uncertainty about the input parameters $\boldsymbol{\Theta}$ via a posterior distribution conditional on the real-world data and uncertainty about the mean simulation response via a posterior distribution on $M(\cdot)$ conditional on a designed simulation experiment; together they provide a posterior distribution and corresponding CrI for U.

The next section describes other Bayesian approaches to input uncertainty. This is followed by a formal description of the problem of interest. In §4, we study a tractable $M/M/\infty$ queue to gain insights about the value of metamodeling, which is key to reducing simulation estimation error. Based on these insights, we introduce a fully Bayesian framework capturing both input and metamodel uncertainty to provide a posterior distribution for U in §5. We then propose a computational procedure to construct the CrI for U. Results from an empirical study of a more practical problem are reported in §6, and we conclude the paper in §7.

2. Background

In the simulation literature various Bayesian approaches for analyzing system performance have been proposed. To facilitate the review, we represent the simulation output on independent replication j when the input parameter is $\boldsymbol{\theta}$ by

$$Y_j(\boldsymbol{\theta}) = \mu(\boldsymbol{\theta}) + \epsilon_j(\boldsymbol{\theta})$$

where $\epsilon_j(\boldsymbol{\theta})$ is a mean-zero, finite-variance random variable representing the output variability of the simulation.

Suppose that $\boldsymbol{\theta}^c$ is known so that we can generate independent and identically distributed (i.i.d.) simulation outputs $\{Y_1(\boldsymbol{\theta}^c), Y_2(\boldsymbol{\theta}^c), \ldots, Y_n(\boldsymbol{\theta}^c)\}$. Andradóttir and Bier (2000) consider a direct application of Bayes' rule to obtain a posterior distribution for $\mathrm{E}[Y(\boldsymbol{\theta}^c)]$ when the distribution family of $Y(\boldsymbol{\theta}^c)$ is assumed known.

Of course, $\boldsymbol{\theta}^c$ is typically unknown but estimable from real-world data. The Bayesian model average (BMA) method proposed by Chick (2001) starts with priors on both the input model families and the values of their parameters. Given real-world data $\mathbf{z_m}$, he constructs posterior distributions, draws B random samples from these posterior distributions, and runs a single simulation replication using each sampled input model. These simulation outputs provide an empirical estimate of the posterior distribution of $Y(\boldsymbol{\Theta})$ given $\mathbf{z_m}$; that is, the predictive distribution of the simulation output given the observed input-model data. This empirical distribution is used to form a point estimate and CI for $\mathrm{E}[Y(\boldsymbol{\Theta}) \mid \mathbf{z_m}] = \mathrm{E}[\mu(\boldsymbol{\Theta}) \mid \mathbf{z_m}]$. *Notice that $E[\mu(\boldsymbol{\Theta}) \mid \mathbf{z_m}]$ depends on the posterior distribution $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m})$ and the particular real-world sample $\mathbf{z_m}$ and is not equal to μ^c in general.* Stated differently, Chick (2001) is interested in a point and interval estimate for the expected simulation response, averaged over the uncertain input parameters, rather than a CrI for the mean response at the true parameters. *Our focus on a posterior distribution of U is a distinguishing feature of this paper.*

The Bayesian simulation-replication algorithms of Zouaoui and Wilson (2003, 2004) also focus on estimation of $\mathrm{E}[\mu(\boldsymbol{\Theta}) \mid \mathbf{z_m}]$ using direct simulation. Zouaoui and Wilson (2003) account for input-parameter uncertainty, similar to this paper, whereas Zouaoui and Wilson (2004) account for both parameter and input-model-family uncertainty, as in Chick (2001). A key difference from Chick (2001) is that Zouaoui and Wilson make multiple simulation replications at each posterior distribution sample of the input models so as to separate the two sources of uncertainty: the input uncertainty and the simulation estimation error. Like Chick (2001), Zouaoui and Wilson (2003) provide CIs for $\mathrm{E}[\mu(\boldsymbol{\Theta}) \mid \mathbf{z_m}]$. When the simulation error is negligible, which would occur if a large number of replications were made at each posterior input-model sample, then their percentile CI for this parameter using a random effects model could be interpreted as an approximation of the perfect fidelity CrI for U. However, as we show later, their CI when interpreted as a CrI is typically wider than necessary.

It is worth noting that in Zouaoui and Wilson (2003) the authors also derive a hierarchical Bayesian framework that could be used to estimate a Bayesian CrI for $\mathrm{E}[\mu(\boldsymbol{\Theta}) \mid \mathbf{z_m}]$ (which, again, is not equal to μ^c in general). This framework is built on a number of homogeneity assumptions, including constant-variance normal distributions for Y and $\mu(\boldsymbol{\Theta})$. *We will provide a valid CrI for U without these conditions.*

Another Bayesian method for simulation output analysis was proposed by Chick (1997). Here the goal is to characterize the posterior distribution of the simulation output as a function of the input model parameters rather than to propagate uncertainty about those parameters to the process mean. Suppose that the distribution of the response Y depends only on its mean (and perhaps some nuisance parameters), and a functional form of the relationship is known, say $\mu(\boldsymbol{\theta}) = g(\boldsymbol{\theta}; \boldsymbol{\beta})$; however, the coefficients $\boldsymbol{\beta}$ are unknown. Let $\mathscr{B}$ denote a random vector that characterizes the uncertainty in our belief about $\boldsymbol{\beta}$. Starting with a prior distribution for $\mathscr{B}$ and simulation outputs $\mathbf{y}_{\mathscr{D}}$ at a collection of input-parameter settings $\boldsymbol{\theta}$, denoted by $\mathscr{D}$, Bayes' rule is used to obtain a posterior distribution $p_{\mathscr{B}}(\boldsymbol{\beta} \mid \mathbf{y}_{\mathscr{D}})$; then for any fixed $\boldsymbol{\theta}$, averaging over the metamodel parameter uncertainty provides a predictive distribution for the simulation response $Y(\boldsymbol{\theta})$. Our approach also characterizes simulation uncertainty using a Bayesian metamodel but with less assumed structure. Also, we combine our Bayesian metamodel with a characterization of uncertainty about $\boldsymbol{\Theta}$ to obtain a measure of uncertainty about U.

In the study of Ng and Chick (2006), the input-parameter uncertainty is approximated by an asymptotic posterior normal distribution, and it is propagated to the output mean via a first-order metamodel in the input parameters. However, this asymptotic approximation is not appropriate when the uncertainty about $\boldsymbol{\Theta}$ is large and $\mu(\cdot)$ is highly nonlinear.

Work outside the stochastic simulation literature that is closely related to ours appears in Oakley and O'Hagan (2002) and Oakley (2004). They consider uncertainty quantification in *deterministic* computer experiments when the values of some parameters are unknown or variable. The uncertainty about these parameters is represented by $\boldsymbol{\Theta} \sim G_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ with $G_{\boldsymbol{\Theta}}(\cdot)$ representing some known distribution for the input distribution parameters. The function $\mu(\cdot)$ itself is unknown and each evaluation is expensive. Their goal is to estimate some property of the distribution of $\mu(\boldsymbol{\Theta})$. The prior belief about $\mu(\cdot)$ is characterized by a stationary Gaussian process (GP), denoted by $M(\cdot)$; see, for instance, Sacks et al. (1989). Then given some simulation outputs $\boldsymbol{\mu}_{\mathscr{D}}$, which denotes the system responses at design points $\mathscr{D}$, Bayes' rule is applied to obtain the posterior distribution $p_M(\cdot \mid \boldsymbol{\mu}_{\mathscr{D}})$ for $M(\cdot)$. This provides a metamodel to propagate the parameter uncertainty to the output response. Thus, there are two sources of uncertainty: parameter and metamodel. Inferences about the impact from these two sources are treated separately in Oakley and O'Hagan (2002) and Oakley (2004). Specifically, they first generate many sample paths from the GP posterior, say, $M^{(i)}(\cdot)$, $i = 1, 2, \ldots, I$. Then to quantify the impact of parameter uncertainty, they compute the response statistic of interest for *each* fixed GP sample path $M^{(i)}(\cdot)$ individually by plugging *all* of the samples from $G_{\boldsymbol{\Theta}}$ into each sample path. Differing from their problem, we characterize uncertainty about the input distribution parameters by a Bayesian approach, our evaluation of $\mu(\boldsymbol{\theta})$ has simulation noise, and we are interested in the combined effect of input-parameter and metamodel uncertainty.

The Bayesian framework introduced in the present paper carries both input and metamodel uncertainty to the output mean estimator. In each case uncertainty is represented by a posterior distribution: the input uncertainty by the posterior $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z}_{\mathbf{m}})$, and the metamodel uncertainty by a GP posterior $p_M(\cdot \mid \mathbf{y}_{\mathscr{D}})$. This combined approach implies a fully Bayesian posterior for U, and based on it we can construct a CrI. Further, the metamodel makes effective use of the simulation budget, so that our CrI is closer to the perfect fidelity CrI $[Q_L, Q_U]$ than an interval obtained by direct simulation.

Our approach completes and extends the prior work. Compared with the Bayesian metamodel approach (BMA) in Chick (1997), we use Bayesian posteriors to characterize both input and metamodel uncertainty without assuming a parametric form of $\mu(\cdot)$. Compared with BMA in Chick (2001), our approach focuses on the posterior and a CrI for U instead of a point estimate of $\mathrm{E}[\mu(\boldsymbol{\Theta}) \mid \mathbf{z}_{\mathbf{m}}]$. Again, μ^c and $\mathrm{E}[\mu(\boldsymbol{\Theta}) \mid \mathbf{z}_{\mathbf{m}}]$ are not the same when there is a finite amount of real-world data and the underlying system mean response is a nonlinear function of the inputs. Compared with the Bayesian simulation-replication algorithm in Zouaoui and Wilson (2003) that also focuses on a point estimate and a CI for $\mathrm{E}[\mu(\boldsymbol{\Theta}) \mid \mathbf{z}_{\mathbf{m}}]$, our framework leads to a fully Bayesian CrI quantifying the overall uncertainty about U while simultaneously reducing the influence of simulation estimation error relative to direct simulation. And compared with the asymptotic approximation in Ng and Chick (2006), our method is appropriate even when the quantity of real-world data is not large. Finally, previous Bayesian treatments of input uncertainty in stochastic simulation do not include the stochastic simulation error in the Bayesian formulation, or if they do then they make strong assumptions.

Perhaps the most important point to make is that our focus is on a Bayesian treatment of $\mu^c = \mu(\boldsymbol{\theta}^c)$, the mean simulation response at the correct input parameter values. We believe that this is the parameter that simulation analysts want: the true mean response independent of their prior distributions or observed data. However, they may well want a Bayesian quantification of U that characterizes the uncertainty in our belief about μ^c using all available information: prior and real-world data on the inputs and prior and simulation data on the response. The framework in this paper provides a provably valid path to attain this objective.

3. Problem Statement and Proposed Approach

Suppose that the stochastic simulation output is a function of random numbers and L independent input distributions $F \equiv \{F_1, F_2, \ldots, F_L\}$. For instance, in the $M/M/\infty$ simulation in §4, $\{F_1, F_2\}$ are the interarrival-time and service-time distributions; in the clinic simulation in §6, $\{F_1, F_2, \ldots, F_6\}$ correspond to interarrival-time distributions, bed occupancy times, and patient class probabilities. To simplify notation, we do not explicitly represent the random numbers that drive the simulation.

The output from the jth independent replication of a simulation with input distribution F can be written as

$$Y_j(F) = \mu(F) + \epsilon_j(F)$$

where $\mu(F)$ denotes the unknown output mean and $\epsilon_j(F)$ represents the simulation error with mean zero. Notice that the simulation output depends on the choice of input distributions. The true "correct" input distributions, denoted by $F^c \equiv \{F_1^c, F_2^c, \ldots, F_L^c\}$, are unknown and are estimated from real-world data. We assume F^c exists.

In this paper, we also assume that the distribution families are known, but not their parameter values. Let an $h_l \times 1$ vector $\boldsymbol{\theta}_l$ denote the parameters for the lth input distribution. By stacking $\boldsymbol{\theta}_l$ with $l = 1, 2, \ldots, L$ together, we have a $d \times 1$ dimensional parameter vector $\boldsymbol{\theta}^\top \equiv (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \ldots, \boldsymbol{\theta}_L^\top)$ with $d \equiv \sum_{l=1}^{L} h_l$. Since the parameters uniquely specify the input models, we can equivalently treat $\mu(\cdot)$ as a function of the input-model parameters. Thus, we rewrite the simulation response as

$$Y_j(\boldsymbol{\theta}) = \mu(\boldsymbol{\theta}) + \epsilon_j(\boldsymbol{\theta}). \tag{2}$$

We assume that the unknown true parameters $\boldsymbol{\theta}^c$ are fixed. However, they are estimated by a random sample of real-world observations. Let m_l denote the number of i.i.d. real-world observations available from the lth input distribution $\mathbf{Z}_{l,m_l} \equiv \{Z_{l,1}, Z_{l,2}, \ldots, Z_{l,m_l}\}$ with $Z_{l,i} \overset{\text{i.i.d}}{\sim} F_l^c$, $i = 1, 2, \ldots, m_l$. Let $\mathbf{Z_m} = \{\mathbf{Z}_{l,m_l}, l = 1, 2, \ldots, L\}$ be the collection of samples from all L input distributions in F^c, where $\mathbf{m} = (m_1, m_2, \ldots, m_L)$. The real-world data are a particular realization of $\mathbf{Z_m}$, denoted $\mathbf{z_m}$. *Given a finite sample of real-world data and a finite simulation budget N, we want to produce a Bayesian CrI for U.*

Standard Bayesian inference about $\boldsymbol{\theta}^c$ represents uncertainty by a random vector $\boldsymbol{\Theta}$ with prior distribution $\pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$. For simplification, we assume $\pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ is such that $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m})$ is a density. After obtaining $\mathbf{z_m}$, our belief is updated by Bayes' rule: the data make some values of the parameters more likely than others and some less likely through weighting by the corresponding likelihood,

$$p_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m}) \propto \pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) \cdot p_{\mathbf{Z_m}}(\mathbf{z_m} \mid \boldsymbol{\theta}),$$

where $p_{\mathbf{Z_m}}$ is the assumed likelihood function of $\mathbf{z_m}$ given the parameters. Thus, uncertainty about the input-model parameters is quantified by the posterior $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m})$. Under some regularity conditions (Section 4.2 in Gelman et al. 2004), the effect of the prior will disappear when we have enough data, but an appropriate prior can reduce the input-parameter uncertainty. Notice that we have abused notation by lumping the parameters, priors, and likelihoods of all L distributions together. Since these distributions are assumed independent they would more naturally be treated individually.

If $\mu(\cdot)$ is known, then the impact of input uncertainty can be characterized by an induced posterior distribution $F_U(\cdot \mid \mathbf{z_m}, \mu(\cdot))$. Further, the uncertainty can be quantified by a two-sided $(1-\alpha)100\%$ CrI $[q_{\alpha/2}(\mathbf{z_m}, \mu(\cdot)), q_{1-\alpha/2}(\mathbf{z_m}, \mu(\cdot))]$, where

$$q_\gamma(\mathbf{z_m}, \mu(\cdot)) \equiv \inf\{q\colon F_U(q \mid \mathbf{z_m}, \mu(\cdot)) \geqslant \gamma\}$$

with $\gamma = \alpha/2, 1-\alpha/2$. In our terminology, this is the perfect fidelity two-sided, equal-tail-probability CrI. When we cannot directly evaluate $F_U(\cdot \mid \mathbf{z_m}, \mu(\cdot))$, we can obtain a Monte Carlo estimate of this CrI, $[\hat{q}_{\alpha/2}(\mathbf{z_m}, \mu(\cdot)), \hat{q}_{1-\alpha/2}(\mathbf{z_m}, \mu(\cdot))]$.

```
1. For $b = 1$ to $B$
   (a) Generate $\boldsymbol{\Theta}_b \sim p_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m})$.
   (b) Compute $\mu_b = \mu(\boldsymbol{\Theta}_b)$.
2. With $\gamma = \alpha/2, 1-\alpha/2$, set $\hat{q}_\gamma(\mathbf{z_m}, \mu(\cdot)) = \mu_{(\lceil B\gamma \rceil)}$
   where $\mu_{(b)}$ denotes the $b$th smallest response in the set $\{\mu_b, b = 1, 2, \ldots, B\}$.
```

Since $\mu(\cdot)$ is typically unknown, a straightforward approach is to use simulation to estimate $\mu(\boldsymbol{\Theta}_b)$. Specifically, in step 1b we could use n simulation replications to estimate $\mu(\boldsymbol{\Theta}_b)$ by $\bar{Y}(\boldsymbol{\Theta}_b) \equiv n^{-1}\sum_{j=1}^n Y_j(\boldsymbol{\Theta}_b)$. Notice that the input processes and the simulation noise are mutually independent. We can then approximate $F_U(\cdot \mid \mathbf{z_m}, \mu(\cdot))$ by

$$F_{\bar{Y}(\boldsymbol{\Theta})}(y \mid \mathbf{z_m}) \equiv \Pr\{\bar{Y}(\boldsymbol{\Theta}) \leqslant y \mid \mathbf{z_m}\}$$

and approximate the perfect fidelity CrI for U by $[\bar{q}_{\alpha/2}(\mathbf{z_m}), \bar{q}_{1-\alpha/2}(\mathbf{z_m})]$ where

$$\bar{q}_\gamma(\mathbf{z_m}) \equiv \inf\{q\colon F_{\bar{Y}(\boldsymbol{\Theta})}(q \mid \mathbf{z_m}) \geqslant \gamma\}$$

with $\gamma = \alpha/2, 1-\alpha/2$. The corresponding Monte Carlo estimate is

$$[\hat{\bar{q}}_{\alpha/2}(\mathbf{z_m}), \hat{\bar{q}}_{1-\alpha/2}(\mathbf{z_m})] \equiv [\bar{Y}_{(\lceil B\alpha/2 \rceil)}, \bar{Y}_{(\lceil B(1-\alpha/2) \rceil)}],$$

where $\bar{Y}_{(b)}$ denotes the bth smallest response in the set $\{\bar{Y}_b = \bar{Y}(\boldsymbol{\Theta}_b), b = 1, 2, \ldots, B\}$. We refer to this as the *direct simulation method*. It is essentially the approach of Chick (2001) and Zouaoui and Wilson (2003).

We estimate the input uncertainty through posterior samples $\{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \ldots, \boldsymbol{\Theta}_B\}$. The order statistics of the estimated responses at these samples are used to estimate the $\alpha/2$ and $1-\alpha/2$ quantiles of the simulation response distribution. To obtain quantile estimates without substantial mean squared errors (MSE), B needs to be large enough that observations in the tails of the distribution are likely when $\alpha = 0.1, 0.05$, and 0.01, the traditional values. A typical recommendation is that B should be at least one thousand. Since at each sample $\boldsymbol{\Theta}_b$ the simulation estimator $\bar{Y}(\boldsymbol{\Theta}_b)$ is more variable than $\mu(\boldsymbol{\Theta}_b)$, we expect $[\hat{\bar{q}}_{\alpha/2}(\mathbf{z_m}), \hat{\bar{q}}_{1-\alpha/2}(\mathbf{z_m})]$ to be stochastically wider than $[\hat{q}_{\alpha/2}(\mathbf{z_m}, \mu(\cdot)), \hat{q}_{1-\alpha/2}(\mathbf{z_m}, \mu(\cdot))]$. Given a tight computational budget, n will be small and the impact from the simulation estimation error could be substantial.

The direct simulation approach ignores any relationship between the mean response at different $\boldsymbol{\theta}$ values. However, a relationship typically exists, and this information can be exploited to make more effective use of the simulation budget. Further, if we treat the unknown response function $\mu(\cdot)$ in a Bayesian manner, then we can obtain a CrI for U that correctly reflects input-parameter and simulation uncertainty; the direct simulation approach does not incorporate the simulation uncertainty into the Bayesian formulation.

We will let a random function $M(\cdot)$ represent our uncertainty about $\mu(\cdot)$. Our prior belief about this function is modeled by a stationary GP prior π_M. Given simulation outputs $\mathbf{y}_{\mathcal{D}}$, the belief is updated to a posterior distribution for $M(\cdot)$, denoted by $p_M(\cdot \mid \mathbf{y}_{\mathcal{D}})$. The computational cost of generating $\mathbf{y}_{\mathcal{D}}$ is N. Notice that since we assume $\mu(\cdot)$ is continuous, both π_M and $p_M(\cdot \mid \mathbf{y}_{\mathcal{D}})$ are measures on the space of continuous functions. Properties of this space depend on the correlation structure of the GP; see Adler (2010, Theorem 3.4.1) for conditions that ensure continuity. Then instead of using direct simulation to estimate the mean response at each sample $\boldsymbol{\Theta}_b \sim p_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m})$, we use $M(\boldsymbol{\Theta}_b)$ to propagate the input uncertainty to the output mean. Our formal posterior on $M(\cdot)$ accounts for metamodel uncertainty, where *metamodel uncertainty* results from a finite amount of simulation effort (design points and replications per design point). Notice that the input processes, simulation noise, and GP are mutually independent.

Within this framework, the posterior distribution for U is

$$\begin{aligned} F_U(u \mid \mathbf{z_m}, \mathbf{y}_{\mathscr{D}}) &\equiv \Pr\{U \leqslant u \mid \mathbf{z_m}, \mathbf{y}_{\mathscr{D}}\} \\ &= \Pr\{M(\boldsymbol{\Theta}) \leqslant u \mid \mathbf{z_m}, \mathbf{y}_{\mathscr{D}}\}. \end{aligned} \tag{3}$$

Based on this posterior, we construct the CrI $[q_{\alpha/2}(\mathbf{z_m}, \mathbf{y}_{\mathscr{D}}), q_{1-\alpha/2}(\mathbf{z_m}, \mathbf{y}_{\mathscr{D}})]$ where

$$q_\gamma(\mathbf{z_m}, \mathbf{y}_{\mathscr{D}}) \equiv \inf\{q\colon F_U(q \mid \mathbf{z_m}, \mathbf{y}_{\mathscr{D}}) \geqslant \gamma\}$$

with $\gamma = \alpha/2, 1-\alpha/2$. Based on B posterior samples, a Monte Carlo estimate of this CrI is $[\hat{q}_{\alpha/2}(\mathbf{z_m}, \mathbf{y}_{\mathscr{D}}), \hat{q}_{1-\alpha/2}(\mathbf{z_m}, \mathbf{y}_{\mathscr{D}})]$. We describe the experiment more precisely in the next section.

Our objective in this paper is to provide a Bayesian framework that quantifies the overall uncertainty about U. Furthermore, given a fixed computational budget, we want to reduce the uncertainty introduced when propagating the input-parameter uncertainty to the output mean. Since $[q_{\alpha/2}(\mathbf{z_m}, \mu(\cdot)),\ q_{1-\alpha/2}(\mathbf{z_m}, \mu(\cdot))]$ is the perfect fidelity two-sided, equal-probability CrI, we want our estimated CrI $[\hat{q}_{\alpha/2}(\mathbf{z_m}, \mathbf{y}_{\mathscr{D}}), \hat{q}_{1-\alpha/2}(\mathbf{z_m}, \mathbf{y}_{\mathscr{D}})]$ to be close to it and closer than what can be obtained with the direct simulation method.

4. Value of Metamodeling

In this section we use a tractable $M/M/\infty$ queue to motivate employing a metamodel instead of direct simulation to propagate input-parameter uncertainty to the output mean. The value of this simple setting is that it clearly illustrates how the benefits from metamodeling arise. Our Bayesian framework to accomplish this more generally is presented in the next section.

Suppose we are interested in estimating the steady-state mean number of customers in an $M/M/\infty$ queue when the unknown true arrival rate is $\theta^c = 1$ and the known mean service time is 5. Thus, the true mean response is $\mu^c = \mu(\theta^c) = 5\theta^c = 5$. In this stylized example each replication of the simulation generates one observation of the number of customers in the queue in steady state, which is Poisson(5θ).

We observe m "real-world" interarrival times $\mathbf{z}_m = \{z_1, z_2, \ldots, z_m\}$, which are actually exponentially distributed with rate θ^c. We know the distribution is exponential but pretend that we do not know θ^c. A noninformative prior is used: $\pi_\Theta(\theta) \propto 1/\theta$. Therefore, the corresponding posterior $p_\Theta(\theta \mid \mathbf{z}_m)$ is Gamma$(m, \sum_{i=1}^m z_i)$ (Ng and Chick 2006). If the response surface function $\mu(\theta) = 5\theta$ were known, then the induced posterior distribution for U would be Gamma$(m, \sum_{i=1}^m z_i/5)$. Thus, given the real-world data, the perfect fidelity posterior distribution $F_U(\cdot \mid \mathbf{z}_m, \mu(\cdot))$ is computable for this simple example.

4.1. Direct Simulation

We first explore using direct simulation to propagate the input uncertainty to the output mean. In this setting, "direct simulation" means the following:

```
1. Observe "real-world" data $\mathbf{z}_m = \{z_1, z_2, \ldots, z_m\}$ i.i.d. Exponential$(\theta^c)$.
2. Form the posterior distribution $p_\Theta(\theta \mid \mathbf{z}_m)$, which is Gamma$(m, \sum_{i=1}^m z_i)$.
3. For $b = 1$ to $B$
    (a) Generate $\Theta_b \sim$ Gamma$(m, \sum_{i=1}^m z_i)$.
    (b) Generate $Y_j(\Theta_b)$, $j = 1, 2, \ldots, n$ that are i.i.d Poisson$(5\Theta_b)$.
    (c) Form $\bar{Y}(\Theta_b) = n^{-1}\sum_{j=1}^n Y_j(\Theta_b)$.
Next $b$
4. Use $\bar{Y}(\Theta_b)$, $b = 1, 2, \ldots, B$, to estimate $F_{\bar{Y}(\Theta)}(\cdot \mid \mathbf{z}_m)$ as an approximation for $F_U(\cdot \mid \mathbf{z}_m, \mu(\cdot))$.
```

Here we will obtain the distribution $F_{\bar{Y}(\Theta)}(\cdot \mid \mathbf{z}_m)$ analytically rather than via step 4. Let $C(\theta) \equiv n\bar{Y}(\theta) = \sum_{j=1}^n Y_j(\theta)$. Since $Y_j(\Theta) \mid \Theta \sim$ Poisson(5Θ), we have $C(\Theta) \mid \Theta \sim$ Poisson$(5n\Theta)$. And we know that $\Theta \mid \mathbf{z}_m \sim$ Gamma$(m, \sum_{i=1}^m z_i)$ is the posterior distribution of Θ given the data. From this we derive the distribution of $\bar{Y}(\Theta)$ when n simulation replications are averaged for each posterior sample from Gamma$(m, \sum_{i=1}^m z_i)$. Using standard methods we can show that

$$C \equiv C(\Theta) \mid \mathbf{z}_m \sim \operatorname{NegBin}\left(m, \frac{\sum_{i=1}^m z_i}{5n + \sum_{i=1}^m z_i}\right).$$

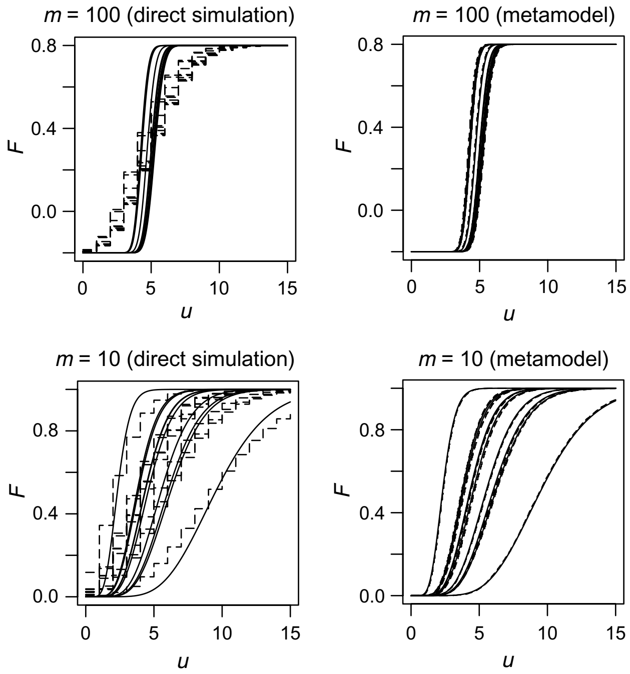
Therefore, $F_{\bar{Y}(\Theta)}(\cdot \mid \mathbf{z}_m)$ is the distribution of C/n, which is computable.

We compare $F_U(\cdot \mid \mathbf{z}_m, \mu(\cdot))$ and $F_{\bar{Y}(\Theta)}(\cdot \mid \mathbf{z}_m)$ in the left panels of Figure 1, where we plot the cumulative distribution functions (cdfs) from 10 macro-replications with real-world sample sizes $m = 10$ and 100, total simulation budget $N = 1{,}000$, and $B = 1{,}000$ samples from the posterior distribution of Θ. Recall that $N = Bn$. In each macro-replication, we first generate m real-world data points $\mathbf{z}_m$, and then conditional on the data, we compute the cdfs rather than do simulation. The solid lines are realizations of $F_U(\cdot \mid \mathbf{z}_m, \mu(\cdot))$, and the dashed lines correspond to $F_{\bar{Y}(\Theta)}(\cdot \mid \mathbf{z}_m)$.

The solid lines are the perfect fidelity posterior distributions, given the real-world data. Their spread indicates the effect of different possible real-world samples. As m increases from 10 to 100, the input uncertainty decreases and $F_U(\cdot \mid \mathbf{z}_m, \mu(\cdot))$ becomes more concentrated around the true response $\mu(\theta^c) = 5$.

Since $F_{\bar{Y}(\Theta)}(\cdot \mid \mathbf{z}_m)$ includes simulation variability, the difference between $F_{\bar{Y}(\Theta)}(\cdot \mid \mathbf{z}_m)$ and $F_U(\cdot \mid \mathbf{z}_m, \mu(\cdot))$ indicates the impact of simulation estimation error. The left two plots in Figure 1 indicate that as m increases from 10 to 100, there is less impact from input uncertainty so that the simulation uncertainty dominates. Notice that as m increases we need even greater simulation effort to remain close to the perfect fidelity posterior distribution for U.

Figure 1. Ten posterior cdfs for each method corresponding to 10 samples of real-world data, where m is the quantity of real-world data in each sample.



Note. Solid lines give the induced posterior $F_U(\cdot\mid\mathbf{z}_m,\mu(\cdot))$; dashed lines give the direct simulation approximation $F_{\bar{Y}(\Theta)}(\cdot\mid\mathbf{z}_m)$ and the metamodel approximation $F_U(\cdot\mid\mathbf{z}_m,\mathbf{y}_{\mathscr{D}})$ with $N=1{,}000$.

4.2. Metamodeling

Assume that we know $\mu(\theta)=\beta\theta$ for the $M/M/\infty$ example but not the value of the slope parameter β.

1. Observe "real-world" data $\mathbf{z}_m=\{z_1,z_2,\ldots,z_m\}$ i.i.d. Exponential(θ^c).

2. Form the posterior distribution $p_\Theta(\theta\mid\mathbf{z}_m)$, which is Gamma$(m,\sum_{i=1}^m z_i)$.

3. Choose a design point θ_0. Expend the entire simulation budget to obtain the outputs $\mathbf{y}_{\mathscr{D}}\equiv\{y_j(\theta_0), j=1,2,\ldots,N\}$. Notice $Y_j(\theta_0)\sim$ Poisson$(\beta\theta_0)$. Without loss of generality, let $\theta_0=1$.

4. For the metamodel parameter β, suppose we have a flat prior $\pi_{\mathscr{B}}(\beta)\propto 1$. By Bayes' rule, the posterior is

$$p_{\mathscr{B}}(\beta\mid\mathbf{y}_{\mathscr{D}})\propto\pi_{\mathscr{B}}(\beta)\cdot p(\mathbf{y}_{\mathscr{D}}\mid\beta)\propto\beta^{\sum_{j=1}^N y_j}e^{-\beta N}.$$

Thus, $\mathscr{B}\mid\mathbf{y}_{\mathscr{D}}\sim$ Gamma$(\sum_{j=1}^N y_j+1,N)$.

```
5. For $b=1$ to $B$
    (a) Generate $\Theta_b\sim \text{Gamma}(m,\sum_{i=1}^m z_i)$.
    (b) Generate $\mathscr{B}_b\sim \text{Gamma}(\sum_{j=1}^N y_j+1,N)$.
    (c) Compute $M_b=\mathscr{B}_b\Theta_b$.
Next $b$
```

6. Use M_b, $b=1,2,\ldots,B$, to approximate $F_U(\cdot\mid\mathbf{z}_m,\mu(\cdot))$.

Again, we will derive the distribution of M_b analytically. Let $U=\mathscr{B}\Theta$. Then we have

$$F_U(u\mid\mathbf{z}_m,\mathbf{y}_{\mathscr{D}})=\int_0^\infty\Pr(\mathscr{B}\theta\leqslant u\mid\mathbf{z}_m,\mathbf{y}_{\mathscr{D}})\cdot\Pr(\Theta=\theta\mid\mathbf{z}_m)\,d\theta$$

$$=\int_0^\infty\Pr(\mathscr{B}\leqslant u/\theta\mid\mathbf{z}_m,\mathbf{y}_{\mathscr{D}})\cdot\frac{(\sum_{i=1}^m z_i)^m\theta^{m-1}e^{-\theta\sum_{i=1}^m z_i}}{(m-1)!}\,d\theta$$

$$=\int_0^\infty\left[1-e^{-(u/\theta)N}\sum_{j=0}^{\sum_{k=1}^N y_k}\frac{1}{j!}\left(\frac{u}{\theta}N\right)^j\right]\cdot\frac{(\sum_{i=1}^m z_i)^m\theta^{m-1}e^{-\theta\sum_{i=1}^m z_i}}{(m-1)!}\,d\theta$$

We compare $F_U(\cdot\mid\mathbf{z}_m,\mu(\cdot))$ and $F_U(\cdot\mid\mathbf{z}_m,\mathbf{y}_{\mathscr{D}})$ in the right panels of Figure 1, where we plot the cdfs from 10 macro-replications, real-world sample sizes $m=10$ and 100, total simulation budget $N=1{,}000$, and $B=1{,}000$ samples from the posterior distribution of Θ. The solid lines are realizations of $F_U(\cdot\mid\mathbf{z}_m,\mu(\cdot))$ and the dashed lines correspond to $F_U(\cdot\mid\mathbf{z}_m,\mathbf{y}_{\mathscr{D}})$. Figure 1 shows that given the same simulation budget, $F_U(\cdot\mid\mathbf{z}_m,\mathbf{y}_{\mathscr{D}})$ is much closer to $F_U(\cdot\mid\mathbf{z}_m,\mu(\cdot))$ than $F_{\bar{Y}(\Theta)}(\cdot\mid\mathbf{z}_m)$ is for either quantity of real-world data. This illustrates the power of using an appropriate metamodel rather than direct simulation.

The example reveals the following insight: Given a finite computational budget, to reduce the impact from the simulation estimation error we should exploit prior information about $\mu(\cdot)$ when we build a metamodel to propagate the input uncertainty to the output mean. Our prior belief about the output mean response surface may be as strong as a global parametric trend or as weak as local smoothness and continuity. As we show in §5, Bayesian metamodeling provides a convenient method to combine different types of information from prior beliefs and simulation results and it also naturally characterizes the metamodel uncertainty.

5. A Bayesian Framework

In this section we introduce a Bayesian framework that provides a posterior distribution for the system mean response U given input-model data and a designed simulation experiment. We also show how to sample from this posterior distribution to obtain a CrI for U. *Thus, if we start with appropriate priors for the input distribution parameters and system mean response surface, our algorithm provides a rigorous Bayesian characterization of the impact from input and simulation uncertainty and a CrI for* U.

5.1. A Bayesian Output Metamodel

In this paper, we focus on cases where the parameters $\boldsymbol{\theta}$ take *continuous* values in open or closed intervals, e.g., location and scale parameters. We assume that the simulation mean response $\mu(\cdot)$ is a continuous function of $\boldsymbol{\theta}$ and model the simulation output Y by

$$[Y_j(\boldsymbol{\theta})\mid\boldsymbol{\Theta}=\boldsymbol{\theta}]=\underbrace{\mathbf{f}(\boldsymbol{\theta})^\top\boldsymbol{\beta}+W(\boldsymbol{\theta})}_{M(\boldsymbol{\theta})}+\epsilon_j(\boldsymbol{\theta}).\tag{4}$$

This model encompasses three sources of uncertainty: input-parameter uncertainty $\boldsymbol{\Theta}$, mean response uncertainty $M(\boldsymbol{\theta})$, and the simulation output uncertainty $\epsilon_j(\boldsymbol{\theta})$. They are assumed mutually independent. We discuss each in turn.

The input-parameter uncertainty begins with a prior distribution $\pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$ for $\boldsymbol{\Theta}$; the uncertainty is reduced by observing real-world data $\mathbf{z_m}$, as represented by the posterior distribution $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m})$.

For the simulation uncertainty we use a normal approximation $\epsilon(\boldsymbol{\theta}) \sim \mathrm{N}(0, \sigma_\epsilon^2(\boldsymbol{\theta}))$. Since the output is often an average of a large number of more basic outputs, this approximation is appropriate for many simulation settings. We are not directly interested in $\sigma_\epsilon^2(\boldsymbol{\theta})$.

Uncertainty about the mean response surface is modeled by a stochastic process $M(\cdot)$, which includes two parts: $\mathbf{f}(\boldsymbol{\theta})^\top\boldsymbol{\beta}$ and $W(\boldsymbol{\theta})$. The trend $\mathbf{f}(\boldsymbol{\theta})^\top\boldsymbol{\beta}$ captures global spatial dependence, where $\mathbf{f}(\boldsymbol{\theta})$ is a $p \times 1$ vector of known basis functions and $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown trend parameters. The first element of $\mathbf{f}(\boldsymbol{\theta})$ is usually 1. If there is no prior information about a parametric trend—which is often the case, including in our empirical study in §6—then we use $\mathbf{f}(\boldsymbol{\theta})^\top\boldsymbol{\beta} = \beta_0$.

Our prior on the remaining local spatial dependence is a mean-zero, second-order stationary GP, denoted by $W(\cdot)$. Specifically, $W(\boldsymbol{\theta}) \sim \mathrm{GP}(0, \tau^2 r(\boldsymbol{\theta}, \boldsymbol{\theta}'))$, where $\tau^2 r(\boldsymbol{\theta}, \boldsymbol{\theta}') = \mathrm{Cov}[W(\boldsymbol{\theta}), W(\boldsymbol{\theta}')]$, so that τ^2 is the marginal process variance and $r(\cdot, \cdot)$ is a correlation function. Based on our previous study (Xie et al. 2010), we use the product-form Gaussian correlation function

$$r(\boldsymbol{\theta}, \boldsymbol{\theta}') = \exp\left(-\sum_{j=1}^{d} \phi_j (\theta_j - \theta_j')^2\right) \tag{5}$$

for the empirical evaluation in §6. Let $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_d)$ be the vector of correlation parameters.

If, in addition, we select the prior for $\boldsymbol{\beta}$ to be Gaussian, $\mathscr{B} \sim \mathrm{N}(\mathbf{b}, \boldsymbol{\Omega})$ with $\mathbf{b}$ and $\boldsymbol{\Omega}$ having appropriate dimensions, then the overall prior uncertainty for $M(\cdot)$ is a GP

$$M(\boldsymbol{\theta}) \sim \mathrm{GP}(\mathbf{f}(\boldsymbol{\theta})^\top\mathbf{b}, \mathbf{f}(\boldsymbol{\theta})^\top\boldsymbol{\Omega}\mathbf{f}(\boldsymbol{\theta}') + \tau^2 r(\boldsymbol{\theta}, \boldsymbol{\theta}'))$$

with parameters $(\tau^2, \boldsymbol{\phi})$ (Rasmussen and Williams 2006). This flexible metamodel provides a convenient way to include various types of prior information about $\mu(\cdot)$: global parametric information can be represented by choosing the basis functions $\mathbf{f}(\boldsymbol{\theta})$ and the prior over $\mathscr{B}$; and local spatial dependence information can be included through the covariance function $\tau^2 r(\cdot, \cdot)$.

To reduce uncertainty about $M(\cdot)$ we choose an experiment design consisting of pairs $\mathscr{D} \equiv \{(\boldsymbol{\theta}_i, n_i), i = 1, 2, \ldots, k\}$ at which to run simulations, where $(\boldsymbol{\theta}_i, n_i)$ denotes the location and the number of replications, respectively, at the ith design point. The simulation outputs at $\mathscr{D}$ are $\mathbf{y}_{\mathscr{D}} \equiv \{(y_1(\boldsymbol{\theta}_i), y_2(\boldsymbol{\theta}_i), \ldots, y_{n_i}(\boldsymbol{\theta}_i)); i = 1, 2, \ldots, k\}$ and the sample mean at design point $\boldsymbol{\theta}_i$ is $\bar{y}(\boldsymbol{\theta}_i) = \sum_{j=1}^{n_i} y_j(\boldsymbol{\theta}_i)/n_i$. Let the sample means at all k design points be $\bar{\mathbf{y}}_{\mathscr{D}} = (\bar{y}(\boldsymbol{\theta}_1), \bar{y}(\boldsymbol{\theta}_2), \ldots, \bar{y}(\boldsymbol{\theta}_k))^T$. Since the use of common random numbers is usually detrimental to prediction (Chen et al. 2012), the outputs at different design points should be independent and the variance of $\bar{\mathbf{y}}_{\mathscr{D}}$ is represented by a $k \times k$ diagonal matrix $\mathbf{C} = \mathrm{diag}\{\sigma_\epsilon^2(\boldsymbol{\theta}_1)/n_1, \sigma_\epsilon^2(\boldsymbol{\theta}_2)/n_2, \ldots, \sigma_\epsilon^2(\boldsymbol{\theta}_k)/n_k\}$.

Given the simulation results at design points $\mathbf{y}_{\mathscr{D}}$, we update our belief about $\mu(\cdot)$. Let $F \equiv (\mathbf{f}(\boldsymbol{\theta}_1), \mathbf{f}(\boldsymbol{\theta}_2), \ldots, \mathbf{f}(\boldsymbol{\theta}_k))$, a $p \times k$ matrix. Let $\boldsymbol{\Sigma}$ be the $k \times k$ local spatial covariance matrix of the design points with $\boldsymbol{\Sigma}_{ij} = \tau^2 r(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j)$ and let $\boldsymbol{\Sigma}(\boldsymbol{\theta}, \cdot)$ be the $k \times 1$ local spatial covariance vector between each design point and a fixed prediction point $\boldsymbol{\theta}$. If the parameters $(\tau^2, \boldsymbol{\phi})$ and $\mathbf{C}$ are known, then the posterior distribution of $M(\cdot)$ is the GP

$$M_p(\boldsymbol{\theta}) \equiv M(\boldsymbol{\theta}) \mid \mathbf{y}_{\mathscr{D}} \sim \mathrm{GP}(m_p(\boldsymbol{\theta}), \sigma_p^2(\boldsymbol{\theta})) \tag{6}$$

where $m_p(\cdot)$ is the minimum MSE linear unbiased predictor

$$m_p(\boldsymbol{\theta}) = \mathbf{f}(\boldsymbol{\theta})^\top\hat{\boldsymbol{\beta}} + \boldsymbol{\Sigma}(\boldsymbol{\theta}, \cdot)^\top(\boldsymbol{\Sigma} + \mathbf{C})^{-1}(\bar{\mathbf{y}}_{\mathscr{D}} - F^\top\hat{\boldsymbol{\beta}}), \tag{7}$$

and the corresponding marginal variance is

$$\begin{aligned}\sigma_p^2(\boldsymbol{\theta}) = {} & \tau^2 - \boldsymbol{\Sigma}(\boldsymbol{\theta}, \cdot)^\top(\boldsymbol{\Sigma} + \mathbf{C})^{-1}\boldsymbol{\Sigma}(\boldsymbol{\theta}, \cdot) \\ & + \eta^\top[\boldsymbol{\Omega}^{-1} + F(\boldsymbol{\Sigma} + \mathbf{C})^{-1}F^\top]^{-1}\eta \end{aligned} \tag{8}$$

where $\hat{\boldsymbol{\beta}} = [\boldsymbol{\Omega}^{-1} + F(\boldsymbol{\Sigma} + \mathbf{C})^{-1}F^\top]^{-1}[F(\boldsymbol{\Sigma} + \mathbf{C})^{-1}\bar{\mathbf{y}}_{\mathscr{D}} + \boldsymbol{\Omega}^{-1}\mathbf{b}]$ and $\eta = \mathbf{f}(\boldsymbol{\theta}) - F(\boldsymbol{\Sigma} + \mathbf{C})^{-1}\boldsymbol{\Sigma}(\boldsymbol{\theta}, \cdot)$ (Rasmussen and Williams 2006). The posterior covariance structure can also be expressed, but it is messy and not needed in our work.

This metamodel includes some commonly used predictors as special cases. If we put a point mass prior on $\tau^2 = 0$, then it becomes a parametric regression model on the space spanned by the basis functions $\mathbf{f}(\cdot)$. If, on the other hand, $\boldsymbol{\Omega}^{-1}$ is a matrix of zeros, which is equivalent to no prior information over the global trend, then the posterior for $M(\cdot)$ becomes the stochastic kriging (SK) metamodel of Ankenman et al. (2010).

By combining the effect of input-parameter and metamodel uncertainty, we can derive the posterior distribution of $U = M(\boldsymbol{\Theta})$. Denote the support of $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m})$ by A. Therefore, conditional on $\mathbf{z_m}$ and $\mathbf{y}_{\mathscr{D}}$, the posterior distribution of U is

$$\begin{aligned} F_U(u \mid \mathbf{z_m}, \mathbf{y}_{\mathscr{D}}) &= \Pr\{U \leqslant u \mid \mathbf{z_m}, \mathbf{y}_{\mathscr{D}}\} \\ &= \int_A \Pr\{M(\boldsymbol{\Theta}) \leqslant u \mid \boldsymbol{\Theta} = \boldsymbol{\theta}, \mathbf{y}_{\mathscr{D}}\} p_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m})\, d\boldsymbol{\theta} \\ &= \int_A \Phi\left(\frac{u - m_p(\boldsymbol{\theta})}{\sigma_p(\boldsymbol{\theta})}\right) p_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m})\, d\boldsymbol{\theta} \end{aligned} \tag{9}$$

where $\Phi(\cdot)$ denotes the cdf of the standard normal distribution.

Since the parameters $(\tau^2, \boldsymbol{\phi})$ and $\mathbf{C}$ are unknown, maximum likelihood estimates are typically used for prediction, and the sample variance is used as an estimate of the simulation variance at design points $\mathbf{C}$; see Ankenman et al. (2010). By inserting these into Equations (7) and (8), we can obtain the estimated mean $\hat{m}_p(\boldsymbol{\theta})$ and variance $\hat{\sigma}_p^2(\boldsymbol{\theta})$. The estimated posterior of $M(\boldsymbol{\theta})$ is Gaussian with mean $\hat{m}_p(\boldsymbol{\theta})$ and variance $\hat{\sigma}_p^2(\boldsymbol{\theta})$. Then by inserting these into Equation (9), we can get the estimated posterior distribution of U. In the next section we sample from this posterior distribution to estimate a two-sided, equal-tail-probability $(1-\alpha)100\%$ CrI $[\hat{q}_{\alpha/2}(\mathbf{z_m}, \mathbf{y}_{\mathscr{D}}), \hat{q}_{1-\alpha/2}(\mathbf{z_m}, \mathbf{y}_{\mathscr{D}})]$ for U.

5.2. Procedure to Construct a CrI

Typically we cannot evaluate (9), but sampling from it is relatively easy:

0. Provide priors on $\boldsymbol{\Theta}$ and $\mathscr{B}$.

1. Identify a design space E of $\boldsymbol{\theta}$ values over which to fit the metamodel. This is done empirically by finding the smallest ellipsoid E that covers a large percentage of random samples from the posterior distribution $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m})$ using the method of Barton et al. (2014). The design space is driven by $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m})$ because the purpose of the metamodel is to map values of $\boldsymbol{\theta}$ into a mean simulation response, and the likelihood of these values is governed by $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m})$. As the amount of real-world data increases, the posterior $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m})$ becomes more concentrated, and therefore E shrinks as it should.

2. To obtain an experiment design $\mathscr{D} = \{(\boldsymbol{\theta}_i, n_i), i = 1, 2, \ldots, k\}$, use a Latin hypercube sample to embed k design points into the design space E and assign equal replications to these points to exhaust N. The choice of k is addressed in Barton et al. (2014), and the use of equal replications is the only sensible allocation in a one-stage design.

3. Run simulations at the design points to obtain outputs $\mathbf{y}_{\mathscr{D}}$. Compute the sample averages $\bar{y}(\boldsymbol{\theta}_i)$ and sample variances $s^2(\boldsymbol{\theta}_i)$ of the simulation outputs, for $i = 1, 2, \ldots, k$. Fit the metamodel to calculate the posterior mean $\hat{m}_p(\boldsymbol{\theta})$ and the variance $\hat{\sigma}_p^2(\boldsymbol{\theta})$ using $\{\bar{y}(\boldsymbol{\theta}_i), s^2(\boldsymbol{\theta}_i), \boldsymbol{\theta}_i,\ i = 1, 2, \ldots, k\}$; see Ankenman et al. (2010).

4. For $b = 1$ to B
   (a) Sample $\boldsymbol{\Theta}_b \sim p_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m})$.
   (b) Sample $M_b \sim \mathrm{N}(\hat{m}_p(\boldsymbol{\Theta}_b), \hat{\sigma}_p^2(\boldsymbol{\Theta}_b))$.

Next b

5. Report an estimated CrI:

$$[\hat{q}_{\alpha/2}(\mathbf{z_m}, \mathbf{y}_{\mathscr{D}}), \hat{q}_{1-\alpha/2}(\mathbf{z_m}, \mathbf{y}_{\mathscr{D}})] \equiv [M_{(\lceil B(\alpha/2)\rceil)}, M_{(\lceil B(1-\alpha/2)\rceil)}] \tag{10}$$

where $M_{(1)} \leqslant M_{(2)} \leqslant \cdots \leqslant M_{(B)}$ are the sorted values.

Step 4 generates B samples $\{M_1, M_2, \ldots, M_B\}$ from the posterior distribution of U according to Equation (9), providing the estimated CrI in (10) whose precision improves as B increases. Beyond the N simulation replications, the additional computational burden depends on how difficult it is to execute step 4a. When we use standard parametric families and conjugate or noninformative priors—as in the next section—sampling from the posteriors is typically fast. Otherwise, we need to resort to some computational approaches such as MCMC to generate samples from $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m})$. Notice that the sampling procedure in step 4 is similar to that used for estimating a conditional expectation in Lee and Glynn (1999) and Steckley and Henderson (2003).

In this paper, we use a Monte Carlo approach to estimate percentiles of the posterior distribution $F_U(\cdot \mid \mathbf{z_m}, \mathbf{y}_{\mathscr{D}})$. Other methods, such as randomized quasi-Monte Carlo, might also be employed for the integration in Equation (9) and could yield smaller error (Lemieux 2009). However, these methods may lose their effectiveness when the dimension of the integral becomes large, as it often will (Caflisch 1998). For example, the critical care facility simulated in §6, a relatively small system, already has $\boldsymbol{\theta}$ with dimension equal to 12. Further, quasi-Monte Carlo is not as versatile as Monte Carlo, and this may be an issue when the posterior $p_{\boldsymbol{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m})$ is not a standard distribution and we need to use computational methods such as MCMC to generate samples from $\boldsymbol{\Theta}$. The combination of quasi-Monte Carlo with MCMC for general situations is still under study; see Caflisch (1998) and Owen and Tribble (2005).

The estimated CrI in Equation (10) characterizes the impact from both input and metamodel uncertainty. If desired, the variance decomposition in Xie et al. (2014) can be used to assess their relative contributions and guide a decision maker as to where to put more effort: If the input uncertainty dominates, then get more real-world data (if possible); if the metamodel uncertainty dominates, then run more simulations; if neither dominates, then both activities are necessary to reduce the overall uncertainty about U.

Theorem 1. *Suppose the parameters* $\boldsymbol{\theta}$ *take continuous values and the simulation mean response surface* $\mu(\cdot)$ *is a continuous function of* $\boldsymbol{\theta}$. *Suppose also that the input processes* Z_{lj}, *the simulation noise* $\epsilon_j(\boldsymbol{\theta})$ *and GP* $M(\boldsymbol{\theta})$ *are mutually independent, and the parameters* $(\tau^2, \boldsymbol{\phi})$ *and* $\mathbf{C}$ *are known. Then given* $\mathbf{z_m}$ *and* $\mathbf{y}_{\mathscr{D}}$,

1. *the posterior distribution for* U *is continuous*;

2. *as* $B \to \infty$, *the empirical distribution based on samples* $\{M_b, b = 1, 2, \ldots, B\}$ *provides a uniformly consistent estimator of the posterior distribution of* U; *and*

3. $\lim_{B\to\infty}[\hat{q}_{\alpha/2}(\mathbf{z_m}, \mathbf{y}_{\mathscr{D}}), \hat{q}_{1-\alpha/2}(\mathbf{z_m}, \mathbf{y}_{\mathscr{D}})] \overset{\text{a.s.}}{=} [q_{\alpha/2}(\mathbf{z_m}, \mathbf{y}_{\mathscr{D}}), q_{1-\alpha/2}(\mathbf{z_m}, \mathbf{y}_{\mathscr{D}})]$.

Proof. Since $F_U(u \mid \mathbf{z_m}, \mathbf{y}_{\mathscr{D}})$ is a weighted sum of normal distributions by Equation (9), the posterior distribution for U is continuous. By the Glivenko-Cantelli Theorem in Van Der Vaart (1998), the empirical distribution of $\{M_1, M_2, \ldots, M_B\}$ with $M_b \overset{\text{i.i.d.}}{\sim} F_U(\cdot \mid \mathbf{z_m}, \mathbf{y}_{\mathscr{D}})$ converges uniformly to $F_U(\cdot \mid \mathbf{z_m}, \mathbf{y}_{\mathscr{D}})$ almost surely (a.s). Since $F_U(u \mid \mathbf{z_m}, \mathbf{y}_{\mathscr{D}})$ is continuous, by applying Lemma 21.2 in Van Der Vaart (1998), as $B \to \infty$ the quantile estimate $\hat{q}_\gamma(\mathbf{z_m}, \mathbf{y}_{\mathscr{D}}) \overset{\text{a.s.}}{\to} q_\gamma(\mathbf{z_m}, \mathbf{y}_{\mathscr{D}})$ for $\gamma = \alpha/2, 1-\alpha/2$. □

Remark: In the result above we assumed that the parameters $(\tau^2, \boldsymbol{\phi}, \mathbf{C})$ are known; this is a common assumption in the kriging literature because including the effect of parameter estimation error makes the posterior distribution of $M(\cdot)$ mathematically and computationally intractable. To apply our method in practice (including the empirical study in §6) we form the plug-in estimators obtained by inserting $\hat{\tau}^2$, $\hat{\boldsymbol{\phi}}$ and $\hat{\mathbf{C}}$ into the relevant expressions. This, too, is common practice.

Ignoring the error in $(\tau^2, \boldsymbol{\phi}, \mathbf{C})$ leaves open the possibility that we could underestimate the metamodel uncertainty. However, based on our experience with SK, this will not be the case, provided we have an adequate experiment design, such as the one developed in Barton et al. (2014) that we use here. A similar observation about parameter insensitivity in the presence of a good experiment design was made by Gano et al. (2006).

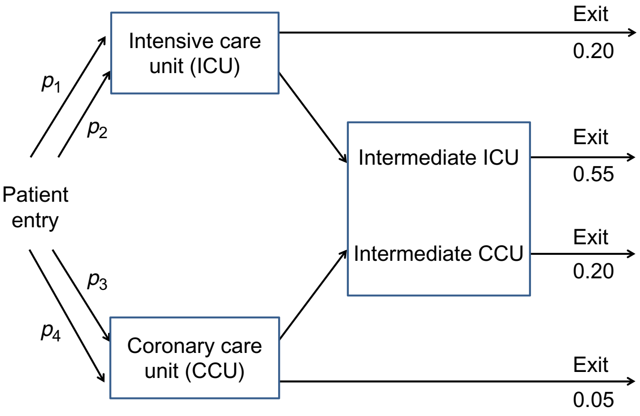
Nevertheless, if one is concerned, then it is possible to apply diagnostic tests such as those described in Bastos and O'Hagan (2009) and Meckesheimer et al. (2002) to evaluate how well the fitted Gaussian process represents the metamodel uncertainty. Yet another approach is to start with prior distributions on the hyperparameters $(\tau^2, \boldsymbol{\phi}, \mathbf{C})$ and thereby include them in the hierarchical Bayesian framework. However, this necessitates a computationally expensive simulation to evaluate the posterior distribution in step 4b. As our results in the next section illustrate, we have not found this to be necessary.

6. Empirical Study

In this section we use the critical care facility described in Ng and Chick (2001) to illustrate the performance of our Bayesian assessment of uncertainty. The structure of the facility is shown in Figure 2. The performance measure is the steady-state expected number of patients per day that are denied entry to the facility. Patients arrive to either the Intensive Care Unit (ICU) or Coronary Care Unit (CCU) and then either exit the facility or go to Intermediate Care (IC), which is a combination of intermediate ICU (IICU) and intermediate CCU (ICCU). Each unit has a finite number of beds. If a patient cannot get an ICU or CCU bed, then he is turned away. If a patient is supposed to move to IC but there is no bed available, then he stays put; when a bed in IC becomes available, the first patient on the waiting list moves in.

The critical care facility includes six input processes. The arrival process is Poisson with arrival rate $\lambda = 3.3$/day (exponentially distributed interarrival times). The stay durations at all four units follow lognormal distributions. Specifically, the ICU stay duration has mean 3.4 days and standard deviation 3.5 days, the CCU stay duration has mean 3.8 days and standard deviation 1.6 days, the IICU stay duration has mean 15.0 days and standard deviation 7.0 days, and the ICCU stay duration has mean 17.0 days and standard deviation 3.0 days. Recall that the density function of the lognormal is

Figure 2. (Color online) Critical care facility.



$$f(x \mid \zeta, \sigma^2) = \frac{1}{x\sqrt{2\pi}\sigma} \exp\left[\frac{-(\ln x - \zeta)^2}{2\sigma^2}\right].$$

There is a one-to-one mapping between input parameters (ζ, σ^2) and the first two moments of the stay time (ζ_L, σ_L^2): mean $\zeta_L = \mathrm{e}^{\zeta + \sigma^2/2}$ and variance $\sigma_L^2 = e^{2\zeta + \sigma^2}(\mathrm{e}^{\sigma^2} - 1)$. The routing probabilities follow a multinomial distribution with parameters $p_1 = 0.2$, $p_2 = 0.55$, $p_3 = 0.2$, and $p_4 = 0.05$. Thus,

$$\begin{aligned}\boldsymbol{\theta} = (&\lambda, \zeta_{\mathrm{ICU}}, \sigma^2_{\mathrm{ICU}}, \zeta_{\mathrm{CCU}}, \sigma^2_{\mathrm{CCU}}, \zeta_{\mathrm{IICU}}, \sigma^2_{\mathrm{IICU}}, \zeta_{\mathrm{ICCU}},\\ &\sigma^2_{\mathrm{ICCU}}, p_1, p_2, p_3, p_4)^\top\end{aligned}$$

and $\boldsymbol{\theta}^c$ is this vector with each element taking the value listed above. Later, when we fit a metamodel for $\mu(\boldsymbol{\theta})$, we drop p_2 since it equals $1 - p_1 - p_3 - p_4$ and is redundant. The number of beds is 14 in ICU, 5 in CCU, and 16 in IC; the IICU and ICCU share the same bed resources.

The goal is to estimate the remaining uncertainty about U, the steady-state expected number of patients per day denied entry, after accounting for all available information: prior and real-world data on the inputs, and prior and simulation data on the response. To evaluate our method, we pretend that the 12 input-model parameters are unknown and estimated from m i.i.d. observations from each of the six true distributions; this represents obtaining "real-world data."

For the interarrival-time process we use the noninformative prior $\pi_\Theta(\lambda) \propto 1/\lambda$. Given m interarrival times $\mathbf{z}_{1,m} = \{z_{1,1}, z_{1,2}, \ldots, z_{1,m}\}$, the posterior distribution $p_\Theta(\lambda \mid \mathbf{z}_{1,m})$ is Gamma$(\psi = m, \delta = \sum_{i=1}^m z_{1,i})$, where ψ and δ denote the shape and rate parameters.

For the stay time at ICU we use a noninformative prior $\pi_{\boldsymbol{\Theta}}(\zeta, \nu) \propto 1/\nu$, where (ζ, ν) denotes the mean and precision $\nu = 1/\sigma^2$ of the logarithm of stay times. Given m real-world observations $\mathbf{z}_{2,m} = \{z_{2,1}, z_{2,2}, \ldots, z_{2,m}\}$, the posterior is

$$p_{\boldsymbol{\Theta}}(\zeta, \nu \mid \mathbf{z}_{2,m}) \propto \underbrace{\frac{(m\nu)^{1/2}}{\sqrt{2\pi}} \exp\left[-\frac{m\nu}{2}(\zeta - \zeta_m)^2\right]}_{p(\zeta|\nu, \mathbf{z}_{2,m}) = \mathrm{N}(\zeta_m, 1/(m\nu))}$$

$$\cdot \underbrace{\frac{\delta_m^{\psi_m}}{\Gamma(\psi_m)}\nu^{\psi_m-1}\exp(-\delta_m\nu)}_{p(\nu|\mathbf{z}_{2,m})=\mathrm{Gamma}(\psi_m,\delta_m)} \qquad (11)$$

where $\zeta_m = \sum_{i=1}^m \ln(z_{2,i})/m$, $\psi_m = (m-1)/2$ and $\delta_m = \sum_{i=1}^m [\ln(z_{2,i}) - \sum_{i=1}^m \ln(z_{2,i})/m]^2/2$ (Ng and Chick 2001). Thus, based on Equation (11), we can generate samples from the posterior $p_{\mathbf{\Theta}}(\zeta, \nu \mid \mathbf{z}_{2,m})$ as follows:

1. Generate ν from $\mathrm{Gamma}(\psi_m, \delta_m)$;
2. Conditional on ν, generate ζ from $\mathrm{N}(\zeta_m, 1/(m\nu))$.

Similarly, we can derive the posterior distributions for the stay-time parameters at the remaining units.

For the routing process, the probabilities p_1, p_2, p_3, and p_4 are estimated from the routing decisions from m patients $\mathbf{z}_{6,m} = \{z_{6,1}, z_{6,2}, \ldots, z_{6,m}\}$. The cumulative numbers of patients choosing the four different routes are denoted by x_1, x_2, x_3, x_4 with $x_j = \sum_{i=1}^m \mathrm{I}(z_{6,i} = j)$ and $\sum_{j=1}^4 x_j = m$, where $\mathrm{I}(\cdot)$ is the indicator function. With a flat prior, the posterior $p_{\mathbf{\Theta}}(p_1, p_2, p_3, p_4 \mid \mathbf{z}_{6,m})$ follows a $\mathrm{Dirichlet}(x_1 + 1, x_2 + 1, x_3 + 1, x_4 + 1)$ distribution (Gelman et al. 2004).

The simulation of the critical care facility starts with an empty system. The first 500 days of startup were discarded as transient (this is sufficient to avoid bias in our study). We consider cases where the computational budget is tight and the simulation estimation uncertainty is significant and cases with low simulation uncertainty. To accomplish the former, we use a short run length for each replication: 10 days after the warm-up. For the latter we use a run length of 500 days after the warm-up.

Ideally we would compare our CrI for U to the perfect fidelity CrI $[q_{\alpha/2}(\mathbf{z_m}, \mu(\cdot)),\ q_{1-\alpha/2}(\mathbf{z_m}, \mu(\cdot))]$, which requires knowledge of $\mu(\cdot)$. Since the true response surface of the critical care facility is not known, we instead used very long simulation runs to estimate the system mean response for each sample $\mathbf{\Theta} \sim p_{\mathbf{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m})$. To find a run length that is adequate to estimate $\mu(\boldsymbol{\theta})$, we did a side experiment: We consider real-world sample sizes of $m = 50, 100, 500$. For each sample size, we ran 10 macro-replications, drawing an independent real-world sample from the true distributions in each. Given these data, we computed the posteriors of the input model parameters; drew 10 samples from each posterior; and recorded estimates of the mean response obtained using run lengths of 10^3, 10^4, 10^5, and 10^6 days. The maximum relative difference for each run length compared to the results obtained using 10^6 is recorded in Table 1. A run length of 10^4 achieved a maximum relative error of 0.05. Considering both the precision and computational cost, we used run length 10^4 days to estimate the system mean response and further to obtain $[\hat{q}_{\alpha/2}(\mathbf{z_m}, \mu(\cdot)), \hat{q}_{1-\alpha/2}(\mathbf{z_m}, \mu(\cdot))]$ for comparison.

We compare our method to direct simulation and to the perfect fidelity CrI. To do so, we ran 1,000 macro-replications of the entire experiment. In each macro-replication, we drew m real-world observations from each input model and computed the posteriors of the input-model parameters $p_{\mathbf{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m})$. To closely approximate the perfect fidelity CrI, we then generated $B = 1{,}000$ posterior samples from $p_{\mathbf{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m})$ and estimated $\mu(\boldsymbol{\theta})$ using a run length of 10^4 days; this yielded $[\hat{q}_{\alpha/2}(\mathbf{z_m}, \mu(\cdot)), \hat{q}_{1-\alpha/2}(\mathbf{z_m}, \mu(\cdot))]$ for that macro-replication.

Table 1. The maximum absolute difference relative to the results obtained by using a run length equal to 10^6 days.

Run length	10^3	10^4	10^5
$m = 50$	0.245	0.05	0.02
$m = 100$	0.191	0.05	0.019
$m = 500$	0.185	0.049	0.016

For direct simulation and our Bayesian approach, we set the run lengths for each simulation replication be 10 or 500 days beyond the warm-up period. A total computational budget of $N = 2{,}000$ replications was expended by each method. For our Bayesian method, the number of design points used to build the metamodel was $k = 20$, implying $n = 100$ replications per design point. For a 12-dimensional problem $k = 20$ is a very small design. We used $B = 1{,}000$ posterior samples to form the CrI. For direct simulation we also used $B = 1{,}000$ posterior samples but allocated $n = 2{,}000/1{,}000 = 2$ replications to each.

The mean and standard deviation (SD) of $\hat{q}^X_{\alpha/2}$, $\hat{q}^X_{1-\alpha/2}$ and the estimated posterior probability content in $[\hat{q}^X_{\alpha/2}, \hat{q}^X_{1-\alpha/2}]$ were obtained for $m = 50{,}100{,}500$ and $\alpha = 0.05$; they are recorded in Tables 2 and 3, where X denotes the method used to obtain the estimate: perfect fidelity, direct simulation, or metamodel. The top halves of Tables 2 and 3 give the results with run length 10 days, implying large simulation estimation uncertainty. The CrI obtained by our Bayesian approach, $[\hat{q}_{\alpha/2}(\mathbf{z_m}, \mathbf{y}_{\mathscr{D}}),\ \hat{q}_{1-\alpha/2}(\mathbf{z_m}, \mathbf{y}_{\mathscr{D}})]$, is very close to $[\hat{q}_{\alpha/2}(\mathbf{z_m}, \mu(\cdot)),\ \hat{q}_{1-\alpha/2}(\mathbf{z_m}, \mu(\cdot))]$. However, as m increases, the difference between direct simulation's $[\hat{\hat{q}}_{\alpha/2}(\mathbf{z_m}),\ \hat{\hat{q}}_{1-\alpha/2}(\mathbf{z_m})]$ and $[\hat{q}_{\alpha/2}(\mathbf{z_m}, \mu(\cdot)),\ \hat{q}_{1-\alpha/2}(\mathbf{z_m}, \mu(\cdot))]$ increases, and the interval obtained by direct simulation is too wide; this is because simulation uncertainty is overwhelming input-parameter uncertainty. On the other hand, since the design space for the GP metamodel is the smallest ellipsoid covering the most likely samples from $p_{\mathbf{\Theta}}(\boldsymbol{\theta} \mid \mathbf{z_m})$, the size of this space decreases as the amount of real-world data m increases. Thus, the metamodel uncertainty decreases. Table 3 shows that as m increases, the error $|\hat{q}_\gamma(\mathbf{z_m}, \mathbf{y}_{\mathscr{D}}) - \hat{q}_\gamma(\mathbf{z_m}, \mu(\cdot))|$ for $\gamma = \alpha/2, 1-\alpha/2$ tends to decrease. Because a larger mean response is typically associated with a larger estimator variance, the estimators of the CrI upper bounds are more variable than they are for the lower bounds.

Since $F_U(\cdot \mid \mathbf{z_m}, \mu(\cdot))$ is unknown, we use the percentage of precisely estimated mean responses contained in the intervals $[\hat{q}_{\alpha/2}(\mathbf{z_m}, \mathbf{y}_{\mathscr{D}}),\ \hat{q}_{1-\alpha/2}(\mathbf{z_m}, \mathbf{y}_{\mathscr{D}})]$ and $[\hat{\hat{q}}_{\alpha/2}(\mathbf{z_m}),\ \hat{\hat{q}}_{1-\alpha/2}(\mathbf{z_m})]$ to estimate the probability content. Tables 2

Table 2. CrI quantile estimates when $m = 50, 100, 500$ and $\alpha = 0.05$, where $\hat{p}^X_\alpha$ denotes the estimated probability content of $F_U(\cdot \mid \mathbf{z_m}, \mu(\cdot))$ in the interval $[\hat{q}^X_{\alpha/2}, \hat{q}^X_{1-\alpha/2}]$.

	$\hat{q}^X_{\alpha/2}$ mean	$\hat{q}^X_{\alpha/2}$ SD	$\hat{q}^X_{1-\alpha/2}$ mean	$\hat{q}^X_{1-\alpha/2}$ SD	$\hat{p}^X_\alpha$ mean	$\hat{p}^X_\alpha$ SD
$m = 50$, run length $= 10$						
Estimated perfect fidelity	1.02	0.4	2.97	0.64		
Direct simulation	0.75	0.37	3.32	0.64	0.99	0.004
GP metamodel	0.94	0.39	2.94	0.64	0.952	0.019
$m = 100$, run length $= 10$						
Estimated perfect fidelity	1.26	0.29	2.62	0.41		
Direct simulation	0.89	0.27	3.05	0.42	0.998	0.002
GP metamodel	1.21	0.29	2.61	0.41	0.951	0.019
$m = 500$, run length $= 10$						
Estimated perfect fidelity	1.63	0.14	2.23	0.16		
Direct simulation	1.06	0.13	2.86	0.18	1	0
GP metamodel	1.61	0.15	2.23	0.17	0.947	0.027
$m = 50$, run length $= 500$						
Estimated perfect fidelity	1	0.39	2.94	0.63		
Direct simulation	0.99	0.39	2.95	0.63	0.951	0.003
GP metamodel	0.93	0.38	2.93	0.63	0.953	0.017
$m = 100$, run length $= 500$						
Estimated perfect fidelity	1.25	0.29	2.61	0.41		
Direct simulation	1.24	0.29	2.62	0.41	0.952	0.004
GP metamodel	1.22	0.29	2.6	0.41	0.95	0.018
$m = 500$, run length $= 500$						
Estimated perfect fidelity	1.63	0.14	2.24	0.17		
Direct simulation	1.61	0.14	2.26	0.17	0.964	0.005
GP metamodel	1.62	0.14	2.23	0.17	0.948	0.017

and 3 show that the probability content of $[\hat{q}_{\alpha/2}(\mathbf{z_m}, \mathbf{y}_\mathcal{D}), \hat{q}_{1-\alpha/2}(\mathbf{z_m}, \mathbf{y}_\mathcal{D})]$ is close to the nominal value $1-\alpha$. However, under the same computational budget, the intervals $[\hat{\hat{q}}_{\alpha/2}(\mathbf{z_m}), \hat{\hat{q}}_{1-\alpha/2}(\mathbf{z_m})]$ obtained by direct simulation are much wider and they typically have obvious over-coverage. Note that "over-coverage" here means the probability content of $F_U(\cdot \mid \mathbf{z_m}, \mu(\cdot))$ contained in the interval is larger than $1-\alpha$; this is different from CI coverage.

Table 3. Errors of CrI quantile estimates when $m = 50$, 100, 500 and $\alpha = 0.05$, where $\mathrm{e}^X_{q_\gamma} \equiv \hat{q}^X_\gamma - \hat{q}_\gamma$ with $\hat{q}_\gamma = \hat{q}_\gamma(\mathbf{z_m}, \mu(\cdot))$ and $\gamma = \alpha/2, 1-\alpha/2$ and $\mathrm{e}^X_{p_\alpha} \equiv \hat{p}^X_\alpha - (1-\alpha)$.

	$\mathrm{e}^X_{q_{\alpha/2}}$ mean	$\mathrm{e}^X_{q_{\alpha/2}}$ SD	$\mathrm{e}^X_{q_{1-\alpha/2}}$ mean	$\mathrm{e}^X_{q_{1-\alpha/2}}$ SD	$\mathrm{e}^X_{p_\alpha}$ mean	$\mathrm{e}^X_{p_\alpha}$ SD
$m = 50$, run length $= 10$						
Direct simulation	−0.27	0.06	0.35	0.07	0.04	0.004
GP metamodel	−0.08	0.09	−0.02	0.12	0.002	0.019
$m = 100$, run length $= 10$						
Direct simulation	−0.37	0.05	0.44	0.06	0.048	0.002
GP metamodel	−0.05	0.06	−0.01	0.08	0.001	0.019
$m = 500$, run length $= 10$						
Direct simulation	−0.57	0.04	0.62	0.04	0.05	0
GP metamodel	−0.02	0.04	0	0.04	−0.003	0.027
$m = 50$, run length $= 500$						
Direct simulation	−0.01	0.02	0.01	0.02	0.001	0.003
GP metamodel	−0.07	0.08	−0.02	0.11	0.003	0.017
$m = 100$, run length $= 500$						
Direct simulation	−0.01	0.02	0.01	0.02	0.002	0.004
GP metamodel	−0.04	0.05	−0.01	0.07	0	0.018
$m = 500$, run length $= 500$						
Direct simulation	−0.02	0.01	0.02	0.01	0.014	0.005
GP metamodel	−0.006	0.02	−0.008	0.03	−0.004	0.017

The over-coverage of $[\hat{\hat{q}}_{\alpha/2}(\mathbf{z_m}),\ \hat{\hat{q}}_{1-\alpha/2}(\mathbf{z_m})]$ becomes even worse when m increases and input uncertainty declines. This indicates that for smaller input uncertainty, we need a larger computational budget for direct simulation so that the impact from the simulation estimation uncertainty becomes negligible.

The bottom halves of Tables 2 and 3 give the results with run length 500 days. The interval $[\hat{\hat{q}}_{\alpha/2}(\mathbf{z_m}),\ \hat{\hat{q}}_{1-\alpha/2}(\mathbf{z_m})]$ is very close to $[\hat{q}_{\alpha/2}(\mathbf{z_m}, \mu(\cdot)),\ \hat{q}_{1-\alpha/2}(\mathbf{z_m}, \mu(\cdot))]$. This indicates that the simulation estimation error is negligible. From these results one might conclude that when the simulation budget is substantial, then direct simulation is slightly better than using a metamodel. However, for consistency we retained a small experiment design of only $k = 20$ design points even with the larger budget and smaller variance outputs; metamodel error would be reduced even further by using more design points.

The finite-sample performance in Tables 2 *and* 3 *demonstrates that when there is a tight computational budget, our Bayesian approach reduces the influence of simulation estimation error and provides a CrI much closer to* $[q_{\alpha/2}(\mathbf{z_m}, \mu(\cdot)), q_{1-\alpha/2}(\mathbf{z_m}, \mu(\cdot))]$ *than does direct simulation*; *when there is sufficient computational budget, both direct simulation and our approach provide CrIs close to* $[q_{\alpha/2}(\mathbf{z_m}, \mu(\cdot)), q_{1-\alpha/2}(\mathbf{z_m}, \mu(\cdot))]$.

7. Conclusions

When we use simulation to evaluate the performance of a stochastic system, there is input and simulation uncertainty in the performance estimates. In this paper, we propose a fully Bayesian framework to quantify the impact from both sources of uncertainty via a CrI for the simulation mean response when evaluated at the true, correct parametric input-model parameters. We do this by propagating the posterior uncertainty about the input-model parameters to the output mean via a GP that characterizes the posterior information about the mean response as a function of the input models given a set of simulation experiments. A flexible metamodel allows us to include various types of prior information about the simulation mean, and this reduces the influence of simulation estimation error. Our Bayesian framework provides a way to sample from the posterior distribution for the system mean response U from which we can produce an asymptotically valid CrI as the number of posterior samples goes to infinity.

An empirical study using a critical care facility demonstrates that when the computational budget is tight, our Bayesian framework makes effective use of the simulation budget and reduces the uncertainty introduced when propagating the input uncertainty to output mean; when there is sufficient computational budget, then both direct simulation and our approach provide intervals that are close to the perfect fidelity CrI. In addition, our approach has good finite-sample performance even when there are several input models including both discrete and continuous distributions.

We have provided a provably valid Bayesian framework to quantify uncertainty in stochastic simulation problems with univariate, independent, parametric input models from known distribution families. Useful extensions of our framework that we will pursue include multivariate input models, input-model-family uncertainty, and nonparametric input models. Some steps in these directions are provided by Biller and Corlu (2011), who developed an approach to quantify the uncertainty of multivariate input models; Chick (2001) and Zouaoui and Wilson (2004), who accounted for both parameter and input-model-family uncertainty; and Song and Nelson (2013), who considered input uncertainty when using the empirical distribution of the real-world data.

Acknowledgments

This paper is based upon work supported by the National Science Foundation under Grant No. CMMI-1068473. The authors thank the Area Editor, Associate Editor, and referees for their helpful guidance.

Endnote

[1] We use $\boldsymbol{\theta}^c$ to denote the unknown true parameters, $\boldsymbol{\Theta}$ to denote a random variable representing our belief about $\boldsymbol{\theta}^c$, and $\boldsymbol{\theta}$ to denote a generic value or function argument.

References

Adler RJ (2010) *The Geometry of Random Fields* (SIAM, Philadelphia).

Andradóttir S, Bier VM (2000) Applying Bayesian ideas in simulation. *Simulation Practice Theory* 8:253–280.

Ankenman BE, Nelson BL, Staum J (2010) Stochastic kriging for simulation metamodeling. *Oper. Res.* 58(2):371–382.

Barton RR (2012) Tutorial: Input uncertainty in output analysis. Laroque C, Himmelspach J, Pasupathy R, Rose O, Uhrmacher AM, eds. *Proc. 2012 Winter Simulation Conf.* (IEEE Computer Society, Washington, DC), 67–78.

Barton RR, Nelson BL, Xie W (2014) Quantifying input uncertainty via simulation confidence intervals. *INFORMS J. Comput.* 26(1):74–87.

Bastos L, O'Hagan A (2009) Diagnostics for Gaussian process emulators. *Technometrics* 51:425–438.

Biller B, Corlu CG (2011) Accounting for parameter uncertainty in large-scale stochastic simulations with correlated inputs. *Oper. Res.* 59(3):661–673.

Caflisch RE (1998) Monte Carlo and quasi-Monte Carlo methods. *Ada Numerica* 7:1–49.

Chen X, Ankenman BE, Nelson BL (2012) The effect of common random numbers on stochastic kriging metamodels. *ACM Trans. Modeling Comput. Simulation* 22(2): Article 7, 1–20.

Chick SE (1997) Bayesian analysis for simulation input and output. Andradóttir S, Healy KJ, Withers DH, Nelson BL, eds. *Proc. 1997 Winter Simulation Conf.* (IEEE Computer Society , Washington, DC), 253–260.

Chick SE (2001) Input distribution selection for simulation experiments: Accounting for input uncertainty. *Oper. Res.* 49(5):744–758.

Gano SE, Renaud JE, Martin JD, Simpson TW (2006) Update strategies for kriging models used in variable fidelity optimization. *Structural Multidisciplinary Optim.* 32:287–298.

Gelman A, Carlin JB, Stern HS, Rubin DB (2004) *Bayesian Data Analysis*, 2nd edn. (Taylor and Francis,, New York).

Lee SH, Glynn PW (1999) Computing the distribution function of a conditional expectation via Monte Carlo: Discrete conditioning spaces. Farrington P, Nembhard H, Sturrock D, Evans G, eds. *Proc. 1999 Winter Simulation Conf.* (IEEE Computer Society , Washington, DC), 1654–1663.

Lemieux C (2009) *Monte Carlo and Quasi-Monte Carlo Sampling* (Springer, New York).

Meckesheimer M, Barton RR, Simpson TW, Booker A (2002) Computationally inexpensive metamodel assessment strategies. *AIAA J.* 40:2053–2060.

Nelson BL (2013) *Foundations and Methods of Stochastic Simulation: A First Course* (Springer, New York).

Ng SH, Chick SE (2001) Reducing input parameter uncertainty for simulations. Peters BA, Smith JS, Medeiros DJ, Rohrer MW, eds. *Proc. 2001 Winter Simulation Conf.* (IEEE Computer Society , Washington, DC), 364–371.

Ng SH, Chick SE (2006) Reducing parameter uncertainty for stochastic systems. *ACM Trans. Modeling Comput. Simulation* 16:26–51.

Oakley J (2004) Estimating percentiles of uncertain computer code outputs. *Appl. Statist.* 53:83–93.

Oakley J, O'Hagan A (2002) Bayesian inference for the uncertainty distribution of computer model outputs. *Biometrika* 89:769–784.

Owen AB, Tribble SD (2005) A quasi-Monte Carlo metropolis algorithm. *Proc. Natl. Acad. Sci. USA* 102:8844–8849.

Rasmussen CE, Williams CKI (2006) *Gaussian Processes for Machine Learning* (MIT Press, London).

Sacks J, Welch WJ, Mitchell TJ, Wynn HP (1989) Design and analysis of computer experiments. *Statist. Sci.* 4:409–435.

Song E, Nelson BL (2013) A quicker assessment of input uncertainty. Pasupathy R, Kim SH, Tolk A, Hill R, Kuhl ME, eds. *Proc. 2013 Winter Simulation Conf.* (IEEE Computer Society , Washington, DC), 474–485.

Steckley SG, Henderson SG (2003) A kernel approach to estimating the density of a conditional expectation. Chick S, Sánchez PJ, Ferrin D, Morrice DJ, eds. *Proc. 2003 Winter Simulation Conf.* (IEEE Computer Society , Washington, DC), 383–391.

Van Der Vaart AW (1998) *Asymptotic Statistics* (Cambridge University Press, Cambridge, UK).

Xie W, Nelson BL, Barton RR (2014) Statistical uncertainty analysis for stochastic simulation. Working Paper, Northwestern University, Evanston, IL.

Xie W, Nelson BL, Staum J (2010) The influence of correlation functions on stochastic kriging metamodels. Johansson B, Jain S, Montoya-Torres J, Hugan J, Yucesan E, eds. *Proc. 2010 Winter Simulation Conf.* (IEEE Computer Society , Washington, DC), 1067–1078.

Zouaoui F, Wilson JR (2003) Accounting for parameter uncertainty in simulation input modeling. *IIE Trans.* 35:781–792.

Zouaoui F, Wilson JR (2004) Accounting for input-model and input-parameter uncertainties in simulation. *IIE Trans.* 36:1135–1151.

Wei Xie is an assistant professor in the department of industrial and systems engineering at Rensselaer Polytechnic Institute. Her research interests are in computer simulation, applied statistics, and data analytics.

Barry L. Nelson is the Walter P. Murphy Professor of Industrial Engineering and Management Sciences at Northwestern University; he is also a Fellow of INFORMS and IIE. His research centers on the design and analysis of computer simulation experiments on models of stochastic systems, and he is the author of *Foundations and Methods of Stochastic Simulation: A First Course*, from Springer.

Russell R. Barton is senior associate dean for research and faculty and professor of supply chain and information systems in the Smeal College of Business, and professor of industrial engineering in the College of Engineering at Penn State. His research interests are in applied statistics, particularly statistical process control and the design and analysis of computer experiments.