Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

METAMODEL-ASSISTED RISK ANALYSIS FOR STOCHASTIC SIMULATION WITH INPUT UNCERTAINTY

Wei Xie, Bo Wang

Industrial and Systems Engineering
Rensselaer Polytechnic Institute
110 8th St.
Troy, NY 12180, USA

Qiong Zhang

Statistical Science and Operations Research
Virginia Commonwealth University
1015 Floyd Ave.
Richmond, VA 23284, USA

ABSTRACT

For complex stochastic systems, simulation can be used to study the system inherent risk behaviors characterized by a sequence of percentiles. In this paper, we develop a Bayesian framework to quantify the overall estimation uncertainty of percentile responses. Suppose that the input parametric families are known. The input model estimation uncertainty is quantified by posterior samples of input parameters. Then, a distributional metamodel is introduced to simultaneously model the percentile response surfaces, which can efficiently propagate the input uncertainty to outputs. Our Bayesian framework can deliver credible intervals for percentiles, and a variance decomposition is further derived to estimate the contributions of input and simulation uncertainties. The empirical studies indicate that our approach has promising performance for system risk analysis.

1 INTRODUCTION

In the current interconnected world, the decision makers often face stochastic systems that are large in scale, complex in behaviors, and have multiple competing objectives (Nelson 2016). For example, there is high uncertainty in the supply, testing, production and demand in the bio-pharmaceutical supply chain risk management. The managers need to make coherent decisions in the procurement, testing and production to improve the system profit, while controlling the impact of uncertainty. *Thus, to construct economic and reliable systems, in this paper, we are interested in the system random behaviors characterized by a sequence of percentiles.* Based on these percentiles, we can estimate other performance measures, such as mean and Conditional Value-at-Risk (CVaR).

Input models are defined as stochastic processes used to drive simulation experiments, for example the inter-arrival and service time distributions of a queueing system, or the product demand distribution of raw materials inventory control. Since the underlying physical input models are unknown and often estimated by finite real-world data, there exists the model estimation error, called *input uncertainty*. Even in the current big data world, we often face the situations where valid real-world data are limited. For example, in the high-tech biopharmaceutical manufacturing, to be competitive, the manufacturers frequently introduce new products, and the product life cycle is short. It takes 9 to 12 months from raw materials sourcing to the final products, and another 2 to 3 months for the quality testing. However, the drug substances could expire after 18 to 36 months, which implies limited demand data available; see Otto et al. (2014). In addition, since each simulation run could be computationally expensive, given limited simulation budget, there exists the simulation estimation uncertainty. Thus, as simulation is used to study the random behaviors of complex stochastic systems, it is necessary to consider both input and simulation uncertainties.

For uncertainty quantification of the system performance estimation, the t-based confidence or credible intervals are not appropriate, and *the percentile intervals* are recommended (Barton 2012). It typically requires large samples of input models to precisely construct a percentile interval quantifying the input

uncertainty. It is computationally prohibitive to propagate the input uncertainty to outputs by using the *direct simulation* which runs simulations at each sample of input models to accurately estimate the system risk behaviors, especially for complex real systems. Compared to the direct simulation, the metamodel constructed based on outputs at a few well-selected design points can efficiently employ the simulation resource and reduce the simulation estimation uncertainty. However, as far as we know, existing metamodel-assisted input uncertainty approaches tend to focus on the system mean response; see for example Cheng and Holland (1997), Cheng and Holland (1998), Cheng and Holland (2004), Xie et al. (2014).

In order to efficiently quantify the impact of input uncertainty on the system random behaviors, we introduce a flexible distributional metamodel that can simultaneously model a sequence of percentiles. There are some recent studies developing metamodel for quantile; see for example Chen and Kim (2014), Wang and Ng (2017), and Batur et al. (2018). Different from those studies which focus on single percentile, we aim at simultaneously developing metamodels for a sequence of percentiles that could be densely selected. It is motivated by the *quantile kriging* in Plumlee and Tuo (2014). However, our methodology is fundamentally different with theirs. First, differing with Plumlee and Tuo (2014) which focus on the system mean response and develop the distributional metamodel for the simulation estimation error, we explore the detailed simulation outputs and construct a rigorous distributional metamodel characterizing the system inherent stochastic uncertainty. Second, Plumlee and Tuo (2014) modeled each percentile curve with kriging without accounting for the fact that the simulation estimation error could change dramatically for different percentiles. Thus, in quantile kriging, the metamodels for different quantile curves share the same GP parameters. In our approach, the Gaussian processes characterizing the simulation estimation uncertainty of percentile curves have different parameters accounting for the fact that the true percentile values and their estimation error depend on the percentile level.

Thus, in the proposed Bayesian framework, we use the posterior distributions of input parameters to quantify the input uncertainty. Then, we explore the detailed simulation outputs and construct *a distributional metamodel* that simultaneously models the response surfaces of a sequence of percentiles. We quantify the simulation uncertainty through the posterior distributions of percentile curves. After that, this metamodel is used to propagate the input uncertainty to outputs, and we can further construct credible intervals (CrIs) quantifying the overall estimation uncertainty of system percentile responses.

In sum, the contributions of our study are described as follows.

- As far as we know, existing studies on metamodel-assisted input uncertainty approaches typically focus on the system mean response. Built on our Bayesian framework introduced in Xie et al. (2014), we study the impact of input uncertainty on a sequence of percentile estimation.
- Differing with the existing approaches on input uncertainty relying on the output summary statistics, we explore the detailed outputs and further develop a distributional metamodel. Our approach can automatically estimate a sequence of percentiles, and it does not require any strong assumption on their spatial dependence structure. In addition, this distributional metamodel can be used to efficiently quantify the impact of input uncertainty on the system random behaviors.
- Our approach can simultaneously deliver CrIs for all percentile responses quantifying the overall estimation uncertainty. A variance decomposition is developed to quantify the contributions from input and simulation uncertainties.

The remaining of this article is organized as follow. Section 2 provides the problem description and the proposed approach. Section 3 introduces the distributional metamodel. A Bayesian framework for percentile uncertainty quantification and a variance decomposition are presented in Section 4. An $M/M/1$ queue example is used to empirically compare the proposed approach with existing stochastic kriging, quantile kriging, and direct simulation methods in Section 5, and we conclude this paper in Section 6.

2 PROBLEM DESCRIPTION AND PROPOSED APPROACH

Simulation is often used to assess the risk behaviors of complex stochastic systems. Suppose that the input distribution families are known, and input models F can be specified by a finite number of parameters,

denoted by $\boldsymbol{\theta}$, with dimension d. At any input parameters $\boldsymbol{\theta}$, the *detailed* simulation outputs are $\mathbf{Y}(\boldsymbol{\theta}) \equiv \{(Y_{r1}(\boldsymbol{\theta}), Y_{r2}(\boldsymbol{\theta}), \ldots, Y_{rL}(\boldsymbol{\theta})), r = 1,2,\ldots,R\}$, where R denotes the number of replications and L denotes the runlength. For example, in the raw materials inventory control, F is the distribution of demand, and Y_{rj} is the overall cost occurring in the j-th ordering time period, including procurement, holding and shortage penalty costs.

In this paper, we focus on quantifying the impact of input uncertainty on the system steady-state random behaviors characterized by a sequence of percentiles. At the underlying "correct" input models, denoted by F^c specified by parameters $\boldsymbol{\theta}^c$, let $Y(\boldsymbol{\theta}^c)$ represent the detailed system output with the distribution, denoted by $G_{Y(\boldsymbol{\theta}^c)}$. We are interested in the random behaviors quantified by a sequence of percentiles, denoted by $\mathbf{q}^c(\boldsymbol{\theta}^c) \equiv (q^c_{\alpha_1}(\boldsymbol{\theta}^c), q^c_{\alpha_2}(\boldsymbol{\theta}^c), \ldots, q^c_{\alpha_\gamma}(\boldsymbol{\theta}^c))$ with

$$q^c_{\alpha_\ell}(\boldsymbol{\theta}^c) \equiv \sup\{q \in \Re : G_{Y(\boldsymbol{\theta}^c)}(q) \leq \alpha_\ell\}$$

for $\ell = 1,2,\ldots,\gamma$, where γ denotes a positive integer and $0 < \alpha_1 < \alpha_2 < \ldots < \alpha_\gamma < 1$ represents the corresponding probabilities. *Thus, we want to estimate the sequence of percentiles* $\mathbf{q}^c(\boldsymbol{\theta}^c)$. To precisely estimate the random behaviors of the detailed output $Y(\boldsymbol{\theta}^c)$, the value of γ could be large. Notice when we consider other system performance measure that is a function of percentiles, the selection of probabilities, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_\gamma)$, could depend on the measure of interests. For example, for the mean response, we could have $\boldsymbol{\alpha}$ evenly covering the range $(0,1)$. If we consider CVaR_β with the right tail probability to be β, the probabilities $\boldsymbol{\alpha}$ could be uniformly distributed on the range $(1-\beta, 1)$.

The simulation outputs depend on the choice of input models. Suppose that the unknown input parameters $\boldsymbol{\theta}^c$ are estimated by m real-world data, denoted by $\mathbf{X}^{(0)}_m \equiv \{X^{(0)}_1, X^{(0)}_2, \ldots, X^{(0)}_m\}$, with $X^{(0)}_i \overset{i.i.d.}{\sim} F^c$ for $i = 1,2,\ldots,m$. The input model estimation uncertainty is quantified by the posterior,

$$p(\boldsymbol{\theta}|\mathbf{X}^{(0)}_m) \propto p(\boldsymbol{\theta}) p(\mathbf{X}^{(0)}_m|\boldsymbol{\theta}),$$

where $p(\boldsymbol{\theta})$ represents the prior distribution and $p(\mathbf{X}^{(0)}_m|\boldsymbol{\theta})$ represents the likelihood of real-world data. Then, we generate B posterior samples of input parameters, denoted by $\widetilde{\boldsymbol{\theta}}^{(b)}$ for $b = 1,2,\ldots,B$, to quantify the input model estimation uncertainty. In this paper, we use $\widetilde{\cdot}$ to represent posterior samples. The posterior samples $\{\mathbf{q}^c(\widetilde{\boldsymbol{\theta}}^{(1)}), \mathbf{q}^c(\widetilde{\boldsymbol{\theta}}^{(2)}), \ldots, \mathbf{q}^c(\widetilde{\boldsymbol{\theta}}^{(B)})\}$ can quantify the impact of input uncertainty on the percentile responses. Let $Q_{\alpha_\ell,b} = q^c_{\alpha_\ell}(\boldsymbol{\theta}^{(b)})$. Thus, if the true percentile response surface $q^c_{\alpha_\ell}(\cdot)$ is known, we can construct a $(1-\xi)\%$ two-sided *fidelity percentile credible interval* (CrI)

$$\text{CrI}^0_{\alpha_\ell} = [Q_{\alpha_\ell,(\lceil(\xi/2)B\rceil)}, Q_{\alpha_\ell,(\lceil(1-\xi/2)B\rceil)}] \tag{1}$$

where $Q_{\alpha_\ell,(1)} \leq Q_{\alpha_\ell,(2)} \leq \ldots \leq Q_{\alpha_\ell,(B)}$ are order statistics of $Q_{\alpha_\ell,b}$ with $b = 1,2,\ldots,B$.

However, the percentile response $q^c_{\alpha_\ell}(\cdot)$ is unknown and estimated by simulation. At any given input parameters $\boldsymbol{\theta}$, we run simulations and explore the *detailed* outputs collected from all R replications,

$$\mathscr{Y}(\boldsymbol{\theta}) \equiv \{\mathscr{Y}_1(\boldsymbol{\theta}), \mathscr{Y}_2(\boldsymbol{\theta}), \ldots, \mathscr{Y}_{RL}(\boldsymbol{\theta})\} = \{Y_{r1}(\boldsymbol{\theta}), \ldots, Y_{rL}(\boldsymbol{\theta}) \text{ with } r = 1,2,\ldots,R\},$$

to efficiently employ the simulation resource and reduce the percentile estimation error. The order statistics of detailed outputs is considered

$$\widehat{q}_{\alpha_\ell}(\boldsymbol{\theta}) = \mathscr{Y}_{(\lceil RL\cdot\alpha_\ell\rceil)}(\boldsymbol{\theta}), \tag{2}$$

for $\ell = 1,2,\ldots,\gamma$, where $\mathscr{Y}_{(1)}(\boldsymbol{\theta}) \leq \mathscr{Y}_{(2)}(\boldsymbol{\theta}) \leq \ldots \leq \mathscr{Y}_{(RL)}(\boldsymbol{\theta})$. It has better asymptotic properties than the classical approach where the percentile estimate is built on the summary statistics of simulation outputs

$$\widehat{q}^S_{\alpha_\ell}(\boldsymbol{\theta}) = \frac{1}{R}\sum_{r=1}^{R} Y_{r,(\lceil L\cdot\alpha_\ell\rceil)}(\boldsymbol{\theta}) \tag{3}$$

where $Y_{r,(\lceil L\cdot\alpha_\ell\rceil)}(\boldsymbol{\theta})$ is the α_ℓ-th order statistics of Y_{rj} for $j = 1,\ldots,L$ with $Y_{r,(1)}(\boldsymbol{\theta}) \leq Y_{r,(2)}(\boldsymbol{\theta}) \leq Y_{r,(L)}(\boldsymbol{\theta})$; see the related asymptotic support in Wang et al. (2018).

To construct precise percentile CrIs quantifying the impact of input uncertainty, the number of posterior samples, B, is large. For complex stochastic systems, it could be computationally prohibitive to precisely estimate $\mathbf{q}^c(\widetilde{\boldsymbol{\theta}}^{(b)})$ for $b = 1, 2, \ldots, B$ since each simulation run could be computationally expensive. Given a finite simulation budget, denoted by C, when the direct simulation is used to propagate the input uncertainty to the output, the number of replications allocated to each sample of input models is very limited, which could lead to high simulation estimation uncertainty, especially for extreme percentiles.

Built on the Bayesian framework introduced in Xie et al. (2014), to efficiently estimate the impact of input uncertainty on the system risk behaviors and reduce the simulation estimation uncertainty, in this paper, *we introduce a distributional metamodel that models the output distribution $G_{Y(\boldsymbol{\theta})}$ as a functional of input parameters $\boldsymbol{\theta}$*. It simultaneously estimates the percentile curves $\mathbf{q}(\boldsymbol{\theta}) = (q_{\alpha_1}(\boldsymbol{\theta}), q_{\alpha_2}(\boldsymbol{\theta}), \ldots, q_{\alpha_\gamma}(\boldsymbol{\theta}))$ with the simulation uncertainty quantified by the posterior distributions of percentile response surfaces. Specifically, given the *detailed* simulation output data, denoted by $\mathscr{Y}_{\mathscr{D}}$, the posterior distribution of $q_{\alpha_\ell}(\cdot)$ is modeled by the GP, denoted by $M_{\alpha_\ell}(\cdot)$, for $\ell = 1, 2, \ldots, \gamma$, where $\mathscr{D}$ denotes the design of experiments. Since the percentiles, $q_{\alpha_1}(\boldsymbol{\theta}), q_{\alpha_2}(\boldsymbol{\theta}), \ldots, q_{\alpha_\gamma}(\boldsymbol{\theta})$, typically increase and decrease simultaneously, we model the sequence of percentile curves, $\mathbf{M}(\cdot) \equiv (M_{\alpha_1}(\cdot), M_{\alpha_2}(\cdot), \ldots, M_{\alpha_\gamma}(\cdot))$, together.

Notice that different from Batur et al. (2018) and Chen and Kim (2014), which focus on single percentile, in our framework, we simultaneously build metamodels for all possible percentiles that could be densely selected. Since all the percentiles depend on the same output data, the corresponding estimators are dependent. Modeling each single percentile curve separately would be inefficient if we ignore this dependence. In addition, it could be challenging to correctly model the dependence between different percentile curves over the design space if we follow the traditional approaches and construct different GP metamodels for percentile curves; see the detailed description in Chen et al. (2013) and Zhou et al. (2011). Our approach can automatically avoid this issue.

Therefore, given the real-world input data $\mathbf{X}_m^{(0)}$ and the simulation outputs $\mathscr{Y}_{\mathscr{D}}$, the posterior distributions of the compound random variables $\mathbf{M}(\boldsymbol{\Theta})$ can quantify the overall estimation uncertainty of $\mathbf{q}^c(\boldsymbol{\theta}^c)$, where $\boldsymbol{\Theta}$ denotes a random vector following the distribution $p(\boldsymbol{\theta}|\mathbf{X}_m^{(0)})$, and $\mathbf{M}(\cdot)$ denoting a multivariate GP quantifies the simulation uncertainty for percentile curves $\mathbf{q}(\cdot)$. Then, through the hierarchical sampling procedure, we construct the CrIs quantifying the overall estimation uncertainty for $\mathbf{q}^c(\boldsymbol{\theta}^c)$ and further develop a variance decomposition to estimate the relative contributions from input and simulation uncertainties.

3 DISTRIBUTIONAL METAMODEL

In this section, we introduce a distributional metamodel to simultaneously model a sequence of percentile curves. As the input parameters are close to each other, the percentile responses tend to be similar. Suppose that the underlying percentile curve $q^c_{\alpha_\ell}(\cdot)$ is a realization of GP. Then, the simulation percentile estimator in Equation (2) can be written as

$$\widehat{q}_{\alpha_\ell}(\boldsymbol{\theta}) = \mu_{\alpha_\ell} + W_{\alpha_\ell}(\boldsymbol{\theta}) + \varepsilon_{\alpha_\ell}(\boldsymbol{\theta})$$

for $\ell = 1, 2, \ldots, \gamma$, where μ_{α_ℓ} is a constant global trend (it can be replaced by a more general trend term $\mathbf{f}(\boldsymbol{\theta})^\top \boldsymbol{\mu}_{\alpha_\ell}$) and $W_{\alpha_\ell}(\boldsymbol{\theta})$ is a zero-mean GP modeling the spatial dependence of the percentile curve $q_{\alpha_\ell}(\boldsymbol{\theta})$. The GP, $M_{\alpha_\ell}(\boldsymbol{\theta}) \equiv \mu_{\alpha_\ell} + W_{\alpha_\ell}(\boldsymbol{\theta})$, characterizes the uncertainty of our belief on the unknown percentile curve $q^c_{\alpha_\ell}(\boldsymbol{\theta})$, and $\varepsilon_{\alpha_\ell}(\boldsymbol{\theta}) \sim \mathscr{N}(0, \sigma^2_{\alpha_\ell}(\boldsymbol{\theta}))$ models the simulation estimation error.

Since the percentile curves with different probabilities typically increase or decrease simultaneously, we model the spatial dependence of percentile curves as

$$\mathrm{Cov}[M_{\alpha_\ell}(\boldsymbol{\theta}_1), M_{\alpha_\ell}(\boldsymbol{\theta}_2)] = \tau^2_{\alpha_\ell} \mathrm{Cor}[Z(\boldsymbol{\theta}_1), Z(\boldsymbol{\theta}_2)]$$

where $Z(\boldsymbol{\theta})$ represents a zero-mean GP with variance equal to one. Thus, the percentile curves share the same correlation parameters. In the empirical study, we use the Gaussian correlation function, $\mathrm{Cor}[Z(\boldsymbol{\theta}_1), Z(\boldsymbol{\theta}_2)] = e^{-\sum_{h=1}^{d} \phi_h (\theta_{1h} - \theta_{2h})^2}$, where $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_d)$ denotes the correlation parameters.

The GP metamodel parameters $\mu_{\alpha_\ell}, \tau^2_{\alpha_\ell}$ and $\sigma^2_{\alpha_\ell}(\boldsymbol{\theta})$ depend on the percentile curve. The spatial variance $\tau^2_{\alpha_\ell}$ and the simulation estimation variance $\sigma^2_{\alpha_\ell}(\boldsymbol{\theta})$ tend to be larger for more extreme percentiles. Suppose that $\sigma^2_\alpha(\boldsymbol{\theta})$ is separable on α and $\boldsymbol{\theta}$, i.e., $\sigma^2_\alpha(\boldsymbol{\theta}) = \sigma^2(\alpha)\cdot\sigma^2(\boldsymbol{\theta})$.

To reduce the estimation uncertainty about the percentile curves, we choose an experiment design $\mathscr{D} \equiv \{(\boldsymbol{\theta}_k, n_k), k = 1,2,\ldots,K\}$ to run simulations and collect detailed output observations, where $(\boldsymbol{\theta}_k, n_k)$ denotes the location and the number of replications at the k-th design point. Then, the percentile estimates $\widehat{\mathbf{q}}^{\mathscr{D}}_{\alpha_\ell} \equiv (\widehat{q}_{\alpha_\ell}(\boldsymbol{\theta}_1), \widehat{q}_{\alpha_\ell}(\boldsymbol{\theta}_2), \ldots, \widehat{q}_{\alpha_\ell}(\boldsymbol{\theta}_K))^\top$ are obtained by using Equation (2) for $\ell = 1,2,\ldots,\gamma$. Without using the common random number across different design points, the variance of $\widehat{\mathbf{q}}^{\mathscr{D}}_{\alpha_\ell}$ is represented by a $(K\times K)$ diagonal matrix, $\Omega_{\alpha_\ell} \equiv \text{diag}\left\{\sigma^2_{\alpha_\ell}(\boldsymbol{\theta}_1), \sigma^2_{\alpha_\ell}(\boldsymbol{\theta}_2), \ldots, \sigma^2_{\alpha_\ell}(\boldsymbol{\theta}_K)\right\}$, where $\sigma^2_{\alpha_\ell}(\boldsymbol{\theta}_k)$ for $k = 1,2,\ldots,K$ is estimated by the bootstrap. Specifically, we draw with replacement from simulation output data $\mathscr{Y}(\boldsymbol{\theta}_k)$ to get the bootstrapped data, denoted by $\mathscr{Y}^{(1)}(\boldsymbol{\theta}_k)$ and calculate the percentile estimate by using Equation (2), denoted by $\widehat{q}^{(1)}_{\alpha_\ell}(\boldsymbol{\theta}_k)$. By repeating this procedure for B_0 times, we get $\widehat{q}^{(1)}_{\alpha_\ell}(\boldsymbol{\theta}_k), \widehat{q}^{(2)}_{\alpha_\ell}(\boldsymbol{\theta}_k), \ldots, \widehat{q}^{(B_0)}_{\alpha_\ell}(\boldsymbol{\theta}_k)$. We plug the sample variance, denoted by $\widehat{\sigma}^2_{\alpha_\ell}(\boldsymbol{\theta}_k)$, into Ω_{α_ℓ} to obtain $\widehat{\Omega}_{\alpha_\ell}$.

For any prediction point $\boldsymbol{\theta}_0$, denote the $(K\times 1)$ spatial covariance vector between $\boldsymbol{\theta}_0$ and design points by $\Sigma_{\alpha_\ell}(\boldsymbol{\theta}_0,\cdot)$, and denote the $(K\times K)$ variance-covariance matrix between design points by Σ_{α_ℓ} for $\ell = 1,2,\ldots,\gamma$. Given the simulation outputs $\mathscr{Y}_{\mathscr{D}}$, the remaining uncertainty of the percentile response surface $q_{\alpha_\ell}(\cdot)$ can be characterized by the updated GP, denoted by $M^p_{\alpha_\ell}(\boldsymbol{\theta}_0) \sim GP(m^p_{\alpha_\ell}(\boldsymbol{\theta}_0), \lambda^2_{\alpha_\ell}(\boldsymbol{\theta}_0))$, with the mean

$$m^p_{\alpha_\ell}(\boldsymbol{\theta}_0) = \widehat{\mu}_{\alpha_\ell} + \Sigma_{\alpha_\ell}(\boldsymbol{\theta}_0,\cdot)(\Sigma_{\alpha_\ell} + \Omega_{\alpha_\ell})^{-1}(\widehat{\mathbf{q}}^{\mathscr{D}}_{\alpha_\ell} - \widehat{\mu}_{\alpha_\ell}\cdot\mathbf{1}_K) \tag{4}$$

and the corresponding variance

$$\lambda^2_{\alpha_\ell}(\boldsymbol{\theta}_0) = \tau^2_{\alpha_\ell} - \Sigma_{\alpha_\ell}(\boldsymbol{\theta}_0,\cdot)^\top(\Sigma_{\alpha_\ell} + \Omega_{\alpha_\ell})^{-1}\Sigma_{\alpha_\ell}(\boldsymbol{\theta}_0,\cdot) + \eta^\top[\mathbf{1}_K^\top(\Sigma_{\alpha_\ell} + \Omega_{\alpha_\ell})^{-1}\mathbf{1}_K]^{-1}\eta \tag{5}$$

where $\widehat{\mu}_{\alpha_\ell} = [\mathbf{1}_K^\top(\Sigma_{\alpha_\ell} + \Omega_{\alpha_\ell})^{-1}\mathbf{1}_K]^{-1}\mathbf{1}_K^\top(\Sigma_{\alpha_\ell} + \Omega_{\alpha_\ell})^{-1}\widehat{\mathbf{q}}^{\mathscr{D}}_{\alpha_\ell}$, $\eta = 1 - \mathbf{1}_K^\top(\Sigma_{\alpha_\ell} + \Omega_{\alpha_\ell})^{-1}\Sigma_{\alpha_\ell}(\boldsymbol{\theta}_0,\cdot)$, and $\mathbf{1}_K$ is a $(K\times 1)$ vector with each element equal to one; see Ankenman et al. (2010) and Yi and Xie (2016).

We plug $\widehat{\Omega}$ into Equations (4) and (5). Then, we need to estimate $\boldsymbol{\phi}$ and $\tau^2_{\alpha_\ell}$ for $\ell = 1,2,\ldots,\gamma$. Since we share the same $\boldsymbol{\phi}$ across different percentile curves, the cross-validation is used to estimate it. Specifically, according to Ankenman et al. (2010), for each percentile response surface, the log-likelihood function of $(\mu_{\alpha_\ell}, \tau^2_{\alpha_\ell}, \boldsymbol{\phi})$ is given by,

$$\begin{aligned} L(\mu_{\alpha_\ell}, \tau^2_{\alpha_\ell}, \boldsymbol{\phi}) = &-\ln[(2\pi)^{K/2}] - \frac{1}{2}\ln[|\tau^2_{\alpha_\ell}R(\boldsymbol{\phi}) + \Omega_{\alpha_\ell}|] \\ &- \frac{1}{2}(\widehat{\mathbf{q}}^{\mathscr{D}}_{\alpha_\ell} - \mu_{\alpha_\ell}\mathbf{1}_K)^\top[\tau^2_{\alpha_\ell}R(\boldsymbol{\phi}) + \Omega_{\alpha_\ell}]^{-1}(\widehat{\mathbf{q}}^{\mathscr{D}}_{\alpha_\ell} - \mu_{\alpha_\ell}\mathbf{1}_K), \end{aligned} \tag{6}$$

where $R(\boldsymbol{\phi})$ is the $(K\times K)$ correlation matrix with the (k_1,k_2)-th entry to be $\text{Cor}[Z(\boldsymbol{\theta}_{k_1}), Z(\boldsymbol{\theta}_{k_2})]$ for $k_1, k_2 \in \{1,2,\cdots,K\}$. After taking partial derivatives with respect to $\tau^2_{\alpha_\ell}$, setting it to zero, and then plugging in $\widehat{\mu}_{\alpha_\ell}$ and $\widehat{\Omega}_{\alpha_\ell}$, we can obtain

$$\text{tr}[\check{\Lambda}^{-1}_{\alpha_\ell}R(\boldsymbol{\phi})] = (\widehat{\mathbf{q}}^{\mathscr{D}}_{\alpha_\ell} - \widehat{\mu}_{\alpha_\ell}\mathbf{1}_K)^\top\check{\Lambda}^{-1}_{\alpha_\ell}R(\boldsymbol{\phi})\check{\Lambda}^{-1}_{\alpha_\ell}(\widehat{\mathbf{q}}^{\mathscr{D}}_{\alpha_\ell} - \widehat{\mu}_{\alpha_\ell}\mathbf{1}_K), \tag{7}$$

where $\check{\Lambda}_{\alpha_\ell} = \tau^2_{\alpha_\ell}R(\boldsymbol{\phi}) + \widehat{\Omega}_{\alpha_\ell}$; see the derivation in Appendix A. Thus, the spatial variance $\tau^2_{\alpha_\ell}$ can be written as a function of $\boldsymbol{\phi}$ for $\ell = 1,2,\ldots,\gamma$. Let $\widehat{\Lambda}_{\alpha_\ell} = \widehat{\tau}^2_{\alpha_\ell}R(\boldsymbol{\phi}) + \widehat{\Omega}_{\alpha_\ell}$. The leave-one-out cross-validation for each α_ℓ and $\boldsymbol{\theta}_i$ can be quickly calculated using,

$$e_{\ell k} = \frac{(\widehat{\Lambda}^{-1}_{\alpha_\ell})_k}{(\widehat{\Lambda}^{-1}_{\alpha_\ell})_{kk}}(\widehat{\mathbf{q}}^{\mathscr{D}}_{\alpha_\ell} - \widehat{\mu}_{\alpha_\ell}\mathbf{1}_K),$$

(see Plumlee and Tuo (2014)), where $()_k$ is the k-th row and $()_{kk}$ is the k-th diagonal element. Thus, we select correlation parameters as,

$$\widehat{\boldsymbol{\phi}} = \arg\min \sum_{\ell=1}^{\gamma}\sum_{k=1}^{K} e^2_{\ell k}.$$

The MLE estimate $\widehat{\tau}^2_{\alpha_\ell}$ can be computed through solving Equation (7) with $\widehat{\boldsymbol{\phi}}$. By plugging in $\widehat{\boldsymbol{\phi}}$ and $\widehat{\tau}^2_{\alpha_\ell}$ for $\ell = 1,2,\ldots,\gamma$ into Equations (4) and (5), we can obtain the GP metamodel for $q_{\alpha_\ell}(\cdot)$.

4 UNCERTAINTY QUANTIFICATION FOR QUANTILE ESTIMATION

In this section, we propose a sampling procedure that can provide a *joint* posterior distribution of percentiles, $\widetilde{\mathbf{Q}} = (\widetilde{Q}_{\alpha_1}, \widetilde{Q}_{\alpha_2}, \ldots, \widetilde{Q}_{\alpha_\gamma})$, where $\widetilde{Q}_{\alpha_\ell} \equiv \widetilde{q}_{\alpha_\ell}(\widetilde{\boldsymbol{\theta}})$ with $\widetilde{\boldsymbol{\theta}} \sim p(\boldsymbol{\theta}|\mathbf{X}^{(0)}_m)$ and $\widetilde{q}_{\alpha_\ell}(\cdot) \sim \text{GP}(m^p_{\alpha_\ell}(\cdot), \lambda^2_{\alpha_\ell}(\cdot))$ for $\ell = 1,2,\ldots,\gamma$. It further delivers percentile CrIs accounting for the overall estimation uncertainty of $\mathbf{q}^c(\boldsymbol{\theta}^c)$. The main steps are shown as follows.

0. Provide the priors on input parameters $\boldsymbol{\Theta}$ and the distributional GP metamodel $\mathbf{M}(\cdot)$.
1. Find the smallest ellipsoid E that covers most likely random samples from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{X}^{(0)}_m)$. To obtain an experiment design $\mathscr{D} = \{(\boldsymbol{\theta}_k, n_k), k = 1,2,\ldots,K\}$, use a Latin hypercube sample to generate K design points evenly covering the design space E, and assign equal replications to these points. See the detailed procedure in Barton et al. (2014).
2. Run simulations at the design points to obtain outputs $\mathscr{Y}_\mathscr{D}$. Construct the distributional metamodel of $M^p_{\alpha_\ell}(\boldsymbol{\theta})$ by using Equations (4) and (5).
3. For $b = 1,2,\ldots,B$,
   (a) Generate a posterior samples of input parameters, $\widetilde{\boldsymbol{\theta}}^{(b)} \sim p(\boldsymbol{\theta}|\mathbf{X}^{(0)}_m)$.
   (b) Generate a sample $(\widetilde{q}_{\alpha_1}(\widetilde{\boldsymbol{\theta}}^{(b)}), \widetilde{q}_{\alpha_2}(\widetilde{\boldsymbol{\theta}}^{(b)}), \ldots, \widetilde{q}_{\alpha_\gamma}(\widetilde{\boldsymbol{\theta}}^{(b)}))$ by using common random number to avoid the crossing issue of different percentiles. Specifically, first generate $U^{(b)} \sim \text{Unif}(0,1)$, and then use the inverse cumulative distribution function (CDF) to generate the α_ℓ-th percentile $\widetilde{q}_{\alpha_\ell}(\widetilde{\boldsymbol{\theta}}^{(b)}) = \Phi^{-1}(U^{(b)}; \widetilde{\boldsymbol{\theta}}^{(b)}, \alpha_\ell)$ with $\ell = 1,2,\ldots,\gamma$, where $\Phi^{-1}(\cdot; \widetilde{\boldsymbol{\theta}}^{(b)}, \alpha_\ell)$ is the inverse CDF of $\mathscr{N}(m^p_{\alpha_\ell}(\widetilde{\boldsymbol{\theta}}^{(b)}), \lambda^2_{\alpha_\ell}(\widetilde{\boldsymbol{\theta}}^{(b)}))$.

   End loop
4. Report the two-sided $(1-\xi)\%$ percentile CrI for $q^c_{\alpha_\ell}(\boldsymbol{\theta}^c)$ with $\ell = 1,2,\ldots,\gamma$

$$\text{CrI}^{DM}_{\alpha_\ell} = [\widetilde{Q}_{\alpha_\ell,(\lceil(\xi/2)B\rceil)}, \widetilde{Q}_{\alpha_\ell,(\lceil(1-\xi/2)B\rceil)}] \tag{8}$$

   where the order statistics $\widetilde{Q}_{\alpha_\ell,(1)} \le \widetilde{Q}_{\alpha_\ell,(2)} \le \ldots \le \widetilde{Q}_{\alpha_\ell,(B)}$ and $\widetilde{Q}_{\alpha_\ell,b} = \widetilde{q}_{\alpha_\ell}(\widetilde{\boldsymbol{\theta}}^{(b)})$ for $b = 1,2,\ldots,B$.

The CrI in Equation (8) constructed by using the distributional metamodel (DM) accounts for the overall uncertainty of $q^c_{\alpha_\ell}(\boldsymbol{\theta}^c)$ with $\ell = 1,2,\ldots,\gamma$. We further conduct the variance decomposition to estimate the relative contributions from input and simulation uncertainties. Given $\mathbf{X}^{(0)}_m$ and $\mathscr{Y}_\mathscr{D}$, the total estimation uncertainty of α_ℓ-th percentile is,

$$\begin{aligned}\text{Var}\left[M_{\alpha_\ell}(\boldsymbol{\Theta})\,\middle|\,\mathbf{X}^{(0)}_m, \mathscr{Y}_\mathscr{D}\right] &= \text{E}_{\boldsymbol{\Theta}}\left[\text{Var}_{M^p_{\alpha_\ell}}(M_{\alpha_\ell}(\boldsymbol{\Theta}) \mid \boldsymbol{\Theta})\,\middle|\,\mathbf{X}^{(0)}_m, \mathscr{Y}_\mathscr{D}\right] + \text{Var}_{\boldsymbol{\Theta}}\left[\text{E}_{M^p_{\alpha_\ell}}(M_{\alpha_\ell}(\boldsymbol{\Theta}) \mid \boldsymbol{\Theta})\,\middle|\,\mathbf{X}^{(0)}_m, \mathscr{Y}_\mathscr{D}\right] \\ &= \text{E}_{\boldsymbol{\Theta}}\left[\lambda^2_{\alpha_\ell}(\boldsymbol{\Theta})\,\middle|\,\mathbf{X}^{(0)}_m, \mathscr{Y}_\mathscr{D}\right] + \text{Var}_{\boldsymbol{\Theta}}\left[m^p_{\alpha_\ell}(\boldsymbol{\Theta})\,\middle|\,\mathbf{X}^{(0)}_m, \mathscr{Y}_\mathscr{D}\right],\end{aligned} \tag{9}$$

where $\sigma^2_M \equiv \text{E}_{\boldsymbol{\Theta}}\left[\lambda^2_{\alpha_\ell}(\boldsymbol{\Theta})|\mathbf{X}^{(0)}_m, \mathscr{Y}_\mathscr{D}\right]$ is a measure of simulation estimation uncertainty and $\sigma^2_I \equiv \text{Var}_{\boldsymbol{\Theta}}\left[m^p_{\alpha_\ell}(\boldsymbol{\Theta})|\mathbf{X}^{(0)}_m, \mathscr{Y}_\mathscr{D}\right]$ is a measure of input uncertainty. We can estimate σ^2_M by sample mean of $\lambda^2_{\alpha_\ell}(\widetilde{\boldsymbol{\theta}}^{(b)})$ and estimate σ^2_I by sample variance $m^p_{\alpha_\ell}(\widetilde{\boldsymbol{\theta}}^{(b)})$ as the following,

$$\widehat{\sigma}^2_M = \frac{1}{B}\sum_{b=1}^{B}\lambda^2_{\alpha_\ell}\left(\widetilde{\boldsymbol{\theta}}^{(b)}\right), \quad \widehat{\sigma}^2_I = \frac{1}{B-1}\sum_{b=1}^{B}\left[m^p_{\alpha_\ell}\left(\widetilde{\boldsymbol{\theta}}^{(b)}\right) - \bar{m}^p_{\alpha_\ell}\right]^2,$$

where $\bar{m}^p_{\alpha_\ell} = \sum_{b=1}^{B} m^p_{\alpha_\ell}(\widetilde{\boldsymbol{\theta}}^{(b)})/B$. Since B is large, we could ignore the finite sampling estimation uncertainty.

5 EMPIRICAL STUDY

In this section, we first compare the performance of our distributional metamodel (DM) with the stochastic kriging (SK) introduced in Ankenman et al. (2010), quantile kriging (QK) introduced by Plumlee and Tuo (2014), and commonly used direct simulation for percentile prediction. Then, an $M/M/1$ queue example is used to study the behaviors of our Bayesian framework for $\mathbf{q}^c(\boldsymbol{\theta}^c)$ estimation uncertainty.

5.1 Prediction of Percentile Response Surfaces

We are interested in the percentiles of time staying in an $M/M/1$ queue with arrival rate equal to one. The unknown input parameter is the utilization, denoted by θ, varying in range $I_\theta = [0.2, 0.95]$. We consider the percentile curves, $q_{\alpha_\ell}(\theta)$ with $\ell = 1,2,\ldots,\gamma$ and $\gamma = 19$, where $(\alpha_1, \alpha_2, \ldots, \alpha_{19}) = (5\%, 10\%, \ldots, 95\%)$. Let the total simulation budget to be $C = 1000$ replications, and let the runlength to be $L = 30, 100$ number of customers.

Given the same simulation budget, we compare the performance of our distributional metamodel (DM), stochastic kriging (SK), quantile kriging (QK), and direct simulation (DS), which is assessed by the percentile prediction at $N_{test} = 500$ test points with $\{\theta_k^{test}, k = 1,2,\ldots,N_{test}\}$ equally spaced on $[0.2, 0.95]$. For the distributional metamodel, by following the "$10d$ rule" (Jones et al. 1998), we select $\mathscr{D} = \{(\theta_k, n_k), k = 1,2,\ldots,10\}$ with design points $\{\theta_1,\ldots,\theta_{10}\}$ equally spaced on $[0.2, 0.95]$, and allocate the same replications at each design point, $n_k = C/10 = 100$. By following the procedure in Section 3, we construct the distributional GP for percentiles $q_{\alpha_\ell}(\theta)$ with $\ell = 1,2,\ldots,\gamma$. For SK, we use the same experiment design $\mathscr{D}$, and at each design point θ_k for $k = 1,2,\ldots,10$, the percentile estimate $\widehat{q}^S_{\alpha_\ell}(\theta_k)$ in Equation (3) is based on the summary order statistics of outputs. The estimation variance at each design point is based on the sample variance of percentile estimation uncertainty over replications, $\frac{1}{n_k(n_k-1)}\sum_{r=1}^{n_k}[Y_{r,(\lceil L\cdot\alpha_\ell\rceil)}(\theta_k) - \bar{Y}_{(\lceil L\cdot\alpha_\ell\rceil)}(\theta_k)]^2$, where $\bar{Y}_{(\lceil L\cdot\alpha_\ell\rceil)}(\theta_k) = \frac{1}{n_k}\sum_{r=1}^{n_k} Y_{r,(\lceil L\cdot\alpha_\ell\rceil)}(\theta_k)$. Then, the SK metamodel introduced in Ankenman et al. (2010) is used to construct the metamodel of $q_{\alpha_\ell}(\cdot)$. In QK, we use the same percentile estimate as in SK. For the direct simulation, we equally allocate the simulation budget to all test points, $n_k^{test} = C/N_{test} = 2$.

Since the steady-state time staying in the $M/M/1$ system follows the exponential distribution with rate equal to service rate minus arrival rate, the true percentiles, $q^c_{\alpha_\ell}(\theta)$ for $\ell = 1,2,\ldots,\gamma$, can be obtained directly. We evaluate the accuracy of estimated percentiles through integrated mean squared error (IMSE) at each percentile level α_ℓ,

$$\text{IMSE}_{\alpha_\ell} = \int_{\theta\in I_\theta} \text{MSE}_{\alpha_\ell}(\theta)d\theta = \int_{\theta\in I_\theta} \left(q^c_{\alpha_\ell}(\theta) - f_{\alpha_\ell}(\theta)\right)^2 d\theta,$$

where $f_{\alpha_\ell}(\cdot)$ is the predicted α_ℓ percentile responses from DM, SK, or DS. We can estimate IMSE,

$$\begin{aligned}\overline{\text{IMSE}}_{\alpha_\ell} &= \frac{1}{W}\sum_{w=1}^{W}\left[\int_{\theta\in I_\theta}\left(q^c_{\alpha_\ell}(\theta) - f^{(w)}_{\alpha_\ell}(\theta)\right)^2 d\theta\right] \\ &\approx \frac{1}{W}\sum_{w=1}^{W}\left\{\sum_{k=2}^{N^{test}}\frac{\theta_k^{test}-\theta_{k-1}^{test}}{2}\left[\left(q^c_{\alpha_\ell}(\theta_k^{test}) - f^{(w)}_{\alpha_\ell}(\theta_k^{test})\right)^2 + \left(q^c_{\alpha_\ell}(\theta_{k-1}^{test}) - f^{(w)}_{\alpha_\ell}(\theta_{k-1}^{test})\right)^2\right]\right\},\end{aligned}$$

where $f^{(w)}_{\alpha_\ell}(\theta_k^{test})$ is the predicted α_ℓ-th percentile at θ_k^{test} from the w-th micro-replication.

The results in Table 1 are obtained by using $W = 200$ macro-replications. At any θ, the underlying true output distribution, $G_{Y(\theta)}$, is exponential and the time staying in the $M/M/1$ system increases dramatically as the utilization approaches one. The upper tail percentiles are much harder to estimate than lower tail, which explains the increasing trend in IMSE as percentile level α_ℓ increases as shown in Table 1. For most cases, the direct simulation performs worst, because given a tight simulation budget, each test point only has few simulation runs. The percentile estimates are very inaccurate, especially for extreme tail part. SK performs slightly better than QK since QK assumes equal estimation variance at different percentile levels, which does not hold. Compared to SK and QK, the distributional metamodel has dominant advantage at

Table 1: IMSE of percentiles estimation of time staying in $M/M/1$ system.

IMSE	$L=30$				$L=100$			
α_ℓ	DM	SK	QK	DS	DM	SK	QK	DS
5%	0.0019	0.3280	0.4583	5.5586	0.0007	0.0716	0.1917	2.9205
10%	0.0059	0.3786	0.4467	5.5506	0.0023	0.1048	0.2014	3.1700
15%	0.0121	0.3654	0.4060	5.4476	0.0047	0.1188	0.1961	3.3293
20%	0.0189	0.3211	0.3456	5.2898	0.0079	0.1173	0.1805	3.4331
25%	0.0262	0.2542	0.2760	5.0809	0.0119	0.1049	0.1590	3.4902
30%	0.0343	0.1789	0.2097	4.8414	0.0170	0.0850	0.1369	3.5207
35%	0.0432	0.1106	0.1596	4.5902	0.0233	0.0650	0.1214	3.5313
40%	0.0517	0.0644	0.1415	4.3357	0.0305	0.0515	0.1217	3.5189
45%	0.0602	0.0599	0.1778	4.1085	0.0388	0.0545	0.1510	3.5004
50%	0.0689	0.1278	0.2989	3.9146	0.0500	0.0933	0.2291	3.4894
55%	0.0837	0.3050	0.5485	3.8076	0.0652	0.1925	0.3858	3.5138
60%	0.1067	0.6535	0.9914	3.8253	0.0854	0.3866	0.6651	3.5984
65%	0.1420	1.2699	1.7317	4.0659	0.1194	0.7372	1.1380	3.8017
70%	0.2166	2.3109	2.9378	4.6564	0.1765	1.3556	1.9256	4.2143
75%	0.3502	4.0400	4.8978	5.8396	0.2794	2.4219	3.2425	4.9963
80%	0.6122	6.9670	8.1580	8.0972	0.4827	4.3100	5.5087	6.4962
85%	1.1789	12.2005	13.8975	12.4324	0.9131	7.8267	9.6573	9.4650
90%	2.5809	22.6219	25.2116	21.5785	1.9018	15.1947	18.2153	15.9533
95%	6.8904	49.5667	54.1472	46.2258	5.0069	35.4785	41.4324	34.5595

most percentile levels, and this advantage becomes even more obvious when we have limited runlength, e.g., $L=30$ case. One reason for the better performance is because the distributional metamodel automatically accounts for the dependence between different percentiles, which could improve the percentile curves estimation. Another reason is that we explore the detailed outputs which can achieve more accurate percentile estimation.

5.2 Percentile Estimation Uncertainty Quantification

In this section, an $M/M/1$ queue is used to compare the performance of proposed Bayesian framework, the SK metamodel assisted approach, and direct simulation on the system risk performance estimation. We are interested in the percentile of time staying in the system, denoted by $\mathbf{q}^c(\boldsymbol{\theta}^c)$. The underlying arrival and service rates $\boldsymbol{\theta}^c=(\theta_1^c,\theta_2^c)=(2,\ 2/0.7)$ are unknown, and they are estimated from finite real-world data $\mathbf{X}_m^{(0)}$. With the non-informative Gamma conjugate prior, the posterior is a Gamma distribution for both arrival or service rates, i.e., $p(\theta_h|\mathbf{X}_m^{(0)})=\Gamma(m,\sum_{i=1}^m X_{hi}^{(0)})$ for $h=1,2$, where $\{X_{hi}^{(0)}: i=1,2,\ldots,m\}$ represents the real-world data for inter-arrival or service times.

We compare the three different methods and construct the 95% two-sided percentile CrI for $q_{\alpha_\ell}(\boldsymbol{\theta})$ with $\ell=1,2,\ldots,\gamma$. For distributional- and SK metamodel-assisted approaches, by following Barton et al. (2014), we find the smallest ellipsoid design space E covering the 99% of posterior samples of input parameters, use Latin hypercube sample to generate $K=20$ design points evenly distributed in the ellipsoid, and assign equal replications to each point. Let the total simulation budget $C=1000$ replications, the runlength $L=30,100$ number of customers, and the size of real-world data $m=300$.

Specifically, we generate $B=1000$ posterior samples $\widetilde{\boldsymbol{\theta}}^{(b)}$ for $b=1,2,\ldots,B$. For our approach, we follow the procedure provided in Section 4, and report the estimated 95% CrI (i.e. $\xi=0.05$). For the SK metamodel assisted approach, we use the same experiment design $\mathscr{D}$, calculate the percentile estimate $\widehat{q}^S_{\alpha_\ell}(\boldsymbol{\theta}_k)$ and its variance at each design point $\boldsymbol{\theta}_k$. Then, we construct the SK metamodel of q_{α_ℓ}, denoted by $GP(\acute{m}^p_{\alpha_\ell}(\cdot),\acute{\lambda}^2_{\alpha_\ell}(\cdot))$; see Ankenman et al. (2010) for the detailed information of SK. To avoid the

percentiles crossing issue, following the same CRN approach described in Section 4, generate samples, $\widetilde{Q}'_{\alpha_\ell,b} = \widetilde{q}'_{\alpha_\ell}(\widetilde{\boldsymbol{\theta}}^{(b)}) = \Phi'^{-1}(U^{(b)};\widetilde{\boldsymbol{\theta}}^{(b)},\alpha_\ell)$ with $\ell = 1,2,\ldots,\gamma$, where $\Phi'^{-1}(\cdot;\widetilde{\boldsymbol{\theta}}^{(b)},\alpha_\ell)$ is the inverse CDF of $\mathscr{N}(\acute{m}^p_{\alpha_\ell}(\widetilde{\boldsymbol{\theta}}^{(b)}),\acute{\lambda}^2_{\alpha_\ell}(\widetilde{\boldsymbol{\theta}}^{(b)}))$. Based on the order statistics $\widetilde{Q}'_{\alpha_\ell,(1)} \leq \widetilde{Q}'_{\alpha_\ell,(2)} \leq \ldots \leq \widetilde{Q}'_{\alpha_\ell,(B)}$, SK gives the two-sided $(1-\xi)\%$ percentile CrI for $q^c_{\alpha_\ell}(\boldsymbol{\theta}^c)$ with $\ell = 1,2,\ldots,\gamma$,

$$\mathrm{CrI}^{SK}_{\alpha_\ell} = [\widetilde{Q}'_{\alpha_\ell,(\lceil(\xi/2)B\rceil)}, \widetilde{Q}'_{\alpha_\ell,(\lceil(1-\xi/2)B\rceil)}]. \tag{10}$$

For the direct simulation, we run stochastic simulation at each posterior sample $\widetilde{\boldsymbol{\theta}}^{(b)}$, and obtain the percentile estimate, denoted by $\widehat{Q}''_{\alpha_\ell,b}$, by using Equation (3), and report the CrI for $q^c_{\alpha_\ell}(\boldsymbol{\theta}^c)$ with $\ell = 1,2,\ldots,\gamma$,

$$\mathrm{CrI}^{DS}_{\alpha_\ell} = [\widehat{Q}''_{\alpha_\ell,(\lceil(\xi/2)B\rceil)}, \widehat{Q}''_{\alpha_\ell,(\lceil(1-\xi/2)B\rceil)}] \tag{11}$$

where $\widehat{Q}''_{\alpha_\ell,(1)} \leq \widehat{Q}''_{\alpha_\ell,(2)} \leq \ldots \leq \widehat{Q}''_{\alpha_\ell,(B)}$ are order statistics.

If the true percentile response surface $q^c_{\alpha_\ell}(\cdot)$ is known, the fidelity or target CrI is

$$\mathrm{CrI}^0_{\alpha_\ell} = [Q_{\alpha_\ell,(\lceil(\xi/2)B\rceil)}, Q_{\alpha_\ell,(\lceil(1-\xi/2)B\rceil)}].$$

We compare the estimated CrI delivered by different approaches to this CrI. We first compute the mean absolute error of lower bound (LB MAE) and upper bound (UB MAE) associated with the CrI obtained from distributional metamodel,

$$\begin{aligned}
\mathrm{MAE}^{DM}_{LB}(\alpha_\ell) &= \frac{1}{W}\sum_{w=1}^{W}\left|Q^{(w)}_{\alpha_\ell,(\lceil(\xi/2)B\rceil)} - \widetilde{Q}^{(w)}_{\alpha_\ell,(\lceil(\xi/2)B\rceil)}\right| \\
\mathrm{MAE}^{DM}_{UB}(\alpha_\ell) &= \frac{1}{W}\sum_{w=1}^{W}\left|Q^{(w)}_{\alpha_\ell,(\lceil(1-\xi/2)B\rceil)} - \widetilde{Q}^{(w)}_{\alpha_\ell,(\lceil(1-\xi/2)B\rceil)}\right|,
\end{aligned} \tag{12}$$

where w indicating results from w-th micro-replication. In our empirical study, we use $W = 200$ macro-replications. By replacing $\widetilde{Q}^{(w)}_{\alpha_\ell,(\lceil(\xi/2)B\rceil)}$, $\widetilde{Q}^{(w)}_{\alpha_\ell,(\lceil(1-\xi/2)B\rceil)}$ in (12) with $\widetilde{Q}'^{(w)}_{\alpha_\ell,(\lceil(\xi/2)B\rceil)}$, $\widetilde{Q}'^{(w)}_{\alpha_\ell,(\lceil(1-\xi/2)B\rceil)}$ or $\widehat{Q}''^{(w)}_{\alpha_\ell,(\lceil(\xi/2)B\rceil)}$, $\widehat{Q}''^{(w)}_{\alpha_\ell,(\lceil(1-\xi/2)B\rceil)}$, we can obtain corresponding lower bound and upper bound MAE for SK and DS methods respectively.

In addition, we consider the coverage probability of proposed CrI at each percentile level α_ℓ, given by $\mathrm{P}\left(q^c_{\alpha_\ell}(\boldsymbol{\Theta}) \in \mathrm{CrI}^{\mathscr{M}}_{\alpha_\ell}\right)$ with $\boldsymbol{\Theta} \sim p(\boldsymbol{\theta}|\mathbf{X}^{(0)}_m)$, where $\mathscr{M} = DM, SK, DS$ implies CrI from different methods (8), (10) and (11), which can be estimated through,

$$\widehat{\mathrm{P}}\left(q^c_{\alpha_\ell}(\widetilde{\boldsymbol{\Theta}}) \in \mathrm{CrI}^{\mathscr{M}}_{\alpha_\ell}\right) = \frac{1}{B}\sum_{b=1}^{B}\mathbb{I}\left(q^c_{\alpha_\ell}(\widetilde{\boldsymbol{\theta}}^{(b)}) \in \mathrm{CrI}^{\mathscr{M}}_{\alpha_\ell}\right)$$

where $\mathbb{I}(\cdot)$ is an indicator function. The CrI width $\left|\mathrm{CrI}^{\mathscr{M}}_{\alpha_\ell}\right|$ is also recorded for evaluating the sharpness of uncertainty quantification. We compute the average coverage probability and CrI width for each α_ℓ percentile obtained from 200 micro-replications as shown in Table 2.

In Table 2, we provide the MAE of estimated CrI lower bound, upper bound, coverage of CrI plus minus corresponding standard error, and the CrI width with standard deviation in parentheses. The direct simulation method has the worst performance. The CrI obtained by our approach has the coverage close to the nominal level 95%. The SK assisted approach tends to have under coverage, and this problem becomes worse at the tail part. Comparing with the SK metamodel assisted and direct simulation approaches, our framework has dominant better performance in most cases.

Our framework can deliver CrIs quantifying the overall uncertainty of percentile estimates. We can further estimate the relative contributions of input and simulation uncertainties. Given a fixed amount of

Table 2: MAE of lower and upper bound, coverage and width of CrI for percentile responses from three methods under input uncertainty.

CrI	Runlength = 30				Runlength = 100			
DM								
α_ℓ	LB MAE	UB MAE	Coverage	CrI Width	LB MAE	UB MAE	Coverage	CrI Width
5%	0.002±0.000	0.010±0.001	0.959±0.002	0.076(0.036)	0.002±0.000	0.009±0.001	0.957±0.001	0.076(0.044)
25%	0.011±0.001	0.049±0.004	0.956±0.001	0.408(0.188)	0.008±0.001	0.050±0.008	0.955±0.001	0.418(0.260)
50%	0.028±0.002	0.114±0.009	0.956±0.001	0.971(0.424)	0.019±0.002	0.104±0.014	0.955±0.001	0.993(0.562)
75%	0.065±0.005	0.234±0.019	0.957±0.001	1.906(0.798)	0.043±0.003	0.231±0.039	0.955±0.001	1.992(1.219)
95%	0.215±0.015	0.712±0.060	0.959±0.002	4.207(1.699)	0.151±0.010	0.572±0.061	0.958±0.001	4.319(2.288)
SK								
α_ℓ	LB MAE	UB MAE	Coverage	CrI Width	LB MAE	UB MAE	Coverage	CrI Width
5%	0.016±0.000	0.151±0.008	0.549±0.008	0.202(0.130)	0.004±0.000	0.059±0.005	0.900±0.002	0.123(0.100)
25%	0.023±0.001	0.175±0.007	0.904±0.003	0.529(0.240)	0.007±0.000	0.106±0.008	0.949±0.001	0.487(0.281)
50%	0.016±0.001	0.081±0.006	0.937±0.001	0.851(0.320)	0.012±0.001	0.070±0.006	0.946±0.001	0.952(0.454)
75%	0.107±0.003	0.765±0.030	0.862±0.002	1.160(0.411)	0.033±0.002	0.415±0.025	0.926±0.001	1.481(0.595)
95%	0.750±0.011	3.290±0.098	0.258±0.006	1.390(0.459)	0.362±0.009	2.424±0.098	0.716±0.004	1.978(0.691)
DS								
α_ℓ	LB MAE	UB MAE	Coverage	CrI Width	LB MAE	UB MAE	Coverage	CrI Width
5%	0.025±0.000	2.141±0.073	1.000±0.000	2.233(1.064)	0.019±0.000	0.479±0.036	1.000±0.000	0.567(0.535)
25%	0.118±0.001	2.567±0.065	1.000±0.000	3.063(1.086)	0.072±0.001	1.140±0.053	1.000±0.000	1.600(0.936)
50%	0.275±0.003	2.650±0.048	1.000±0.000	3.835(1.063)	0.164±0.002	1.717±0.052	0.999±0.000	2.816(1.176)
75%	0.579±0.007	2.147±0.027	0.998±0.000	4.544(1.040)	0.360±0.005	1.928±0.033	0.998±0.000	4.158(1.251)
95%	1.461±0.016	0.702±0.052	0.970±0.001	5.095(1.005)	1.021±0.015	0.865±0.049	0.983±0.001	5.423(1.227)

real-world data with $m = 1000$, the variance decomposition for four cases with $L = 30, 50$ and $C = 300, 800$ is provided in Figure 1. The solid line represents the CDF of $q^c_{\alpha_\ell}(\boldsymbol{\Theta})$ with $\boldsymbol{\Theta} \sim p(\boldsymbol{\theta}|\mathbf{X}^{(0)}_m)$. The segments denote the CrIs for $10\%, 20\%, \ldots, 90\%$ percentiles obtained by our framework. The orange and blue proportions are corresponding variance contribution from input and simulation uncertainties. The overall uncertainty increases as the level of percentile (α_ℓ) increases. Given a fixed amount of real-world data, the overall estimation uncertainty and the proportion of simulation uncertainty decreases as either runlength or replications increases. As we increase the runlength to 50 and the total number of replications to 800, we can accurately estimate the impact of input uncertainty on the system random behaviors since the simulation estimation error only contributes about 5% of overall uncertainty, which indicates that the additional simulation runs are not needed.

6 CONCLUSIONS

We consider the system random behaviors characterized by a sequence of percentiles. We develop a Bayesian framework quantifying the overall estimation uncertainty of percentile responses. The input uncertainty is quantified by posterior distributions of input parameters. Further, a distributional metamodel is introduced to explore the detailed outputs and simultaneously model percentile response surfaces. It can efficiently propagate the input uncertainty to outputs. Our framework can deliver credible intervals of percentiles, and further quantify the relative contributions from input and simulation uncertainties. The empirical study demonstrates the our approach could efficiently utilize the simulation resource for system risk analysis.

A APPENDICES

Here, we derive Equation (7). For the log-likelihood in (6), take the partial derivatives with respect to $\tau^2_{\alpha_\ell}$, and apply results (equations (22) and (23)) in Ankenman et al. (2010),

$$\frac{\partial L(\mu_{\alpha_\ell}, \tau^2_{\alpha_\ell}, \boldsymbol{\phi})}{\partial \tau^2_{\alpha_\ell}} = -\frac{1}{2}\mathrm{tr}\left[(\tau^2_{\alpha_\ell} R(\boldsymbol{\phi}) + \Omega_{\alpha_\ell})^{-1} R(\boldsymbol{\phi})\right] + \frac{1}{2}(\widehat{\mathbf{q}}^{\mathscr{D}}_{\alpha_\ell} - \mu_{\alpha_\ell}\mathbf{1}_K)^\top [\tau^2_{\alpha_\ell} R(\boldsymbol{\phi}) + \Omega_{\alpha_\ell}]^{-1} R(\boldsymbol{\phi}) [\tau^2_{\alpha_\ell} R(\boldsymbol{\phi}) + \Omega_{\alpha_\ell}]^{-1} (\widehat{\mathbf{q}}^{\mathscr{D}}_{\alpha_\ell} - \mu_{\alpha_\ell}\mathbf{1}_K).$$

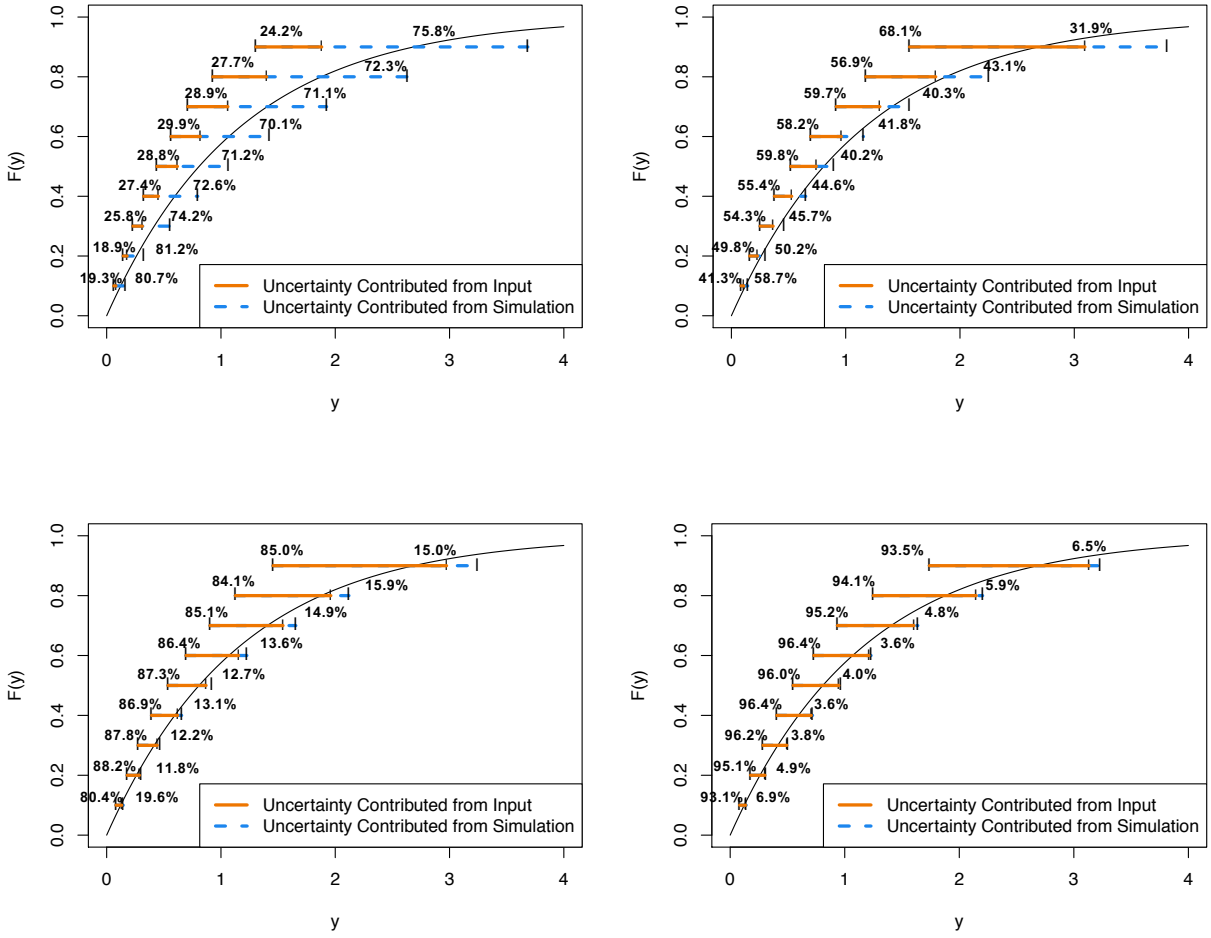

Figure 1: Variance decomposition for percentile responses of $M/M/1$ total time with utilization 0.7. Top-left panel: $L = 30$ and $C = 300$; top-right panel: $L = 50$ and $C = 300$; bottom-left panel: $L = 30$ and $C = 800$; bottom-right panel: $L = 50$ and $C = 800$.

By setting the partial derivative to be zero and plugging in estimated $\widehat{\mu}_{\alpha_\ell}$ and $\widehat{\Omega}_{\alpha_\ell}$, Equation (7) follows.

REFERENCES

Ankenman, B. E., B. L. Nelson, and J. Staum. 2010. "Stochastic kriging for simulation metamodeling". *Operations Research* 58:371–382.

Barton, R. R. 2012. "Tutorial: input uncertainty in output analysis". In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. L. et al., 67–78. Piscataway, New Jersey: IEEE.

Barton, R. R., B. L. Nelson, and W. Xie. 2014. "Quantifying input uncertainty via simulation confidence intervals". *Informs Journal on Computing* 26:74–87.

Batur, D., J. M. Bekki, and X. Chen. 2018. "Quantile regression metamodeling: toward improved responsiveness in the high-tech electronic manufacturing industry". *European Journal of Operational Research* 264:212–224.

Chen, X., and K.-K. Kim. 2014. "Stochastic kriging with biased sample estimates". *ACM Transactions on Modeling and Computer Simulation* 24(2):8.

Chen, X., K. Wang, and F. Yang. 2013. "Stochastic kriging with qualitative factors". In *Proceedings of the 2013 Winter Simulation Conference*, edited by R. P. et al., 790–801. Piscataway, New Jersey: IEEE.

Cheng, R. C. H., and W. Holland. 1997. "Sensitivity of computer simulation experiments to errors in input data". *Journal of Statistical Computation and Simulation* 57:219–241.

Cheng, R. C. H., and W. Holland. 1998. "Two-point methods for assessing variability in simulation output.". *Journal of Statistical Computation and Simulation* 60:183–205.

Cheng, R. C. H., and W. Holland. 2004. "Calculation of confidence intervals for simulation output". *ACM Transactions on Modeling and Computer Simulation* 14:344–362.

Jones, D., M. Schonlau, and W. Welch. 1998. "Efficient global optimization of expensive black-box functions". *Journal of Global Optimization* 13:455–492.

Nelson, B. L. 2016. "'Some tactical problems in digital simulation' for the next 10 years". *Journal of Simulation* 10. 2–11.

Otto, R., A. Santagostino, and U. Schrader. 2014. *From Science to Operations: Questions, Choices and Strategies for Success in Biopharma*. Oxford: Claredon.

Plumlee, M., and R. Tuo. 2014. "Building accurate emulators for stochastic simulations via quantile kriging". *Technometrics* 56:465–473.

Wang, B., Q. Zhang, and W. Xie. 2018. "The bahadur representation of sample quantile for mixing detailed sample path data from stochastic simulation". Working paper.

Wang, S., and S. H. Ng. 2017. "A joint Gaussian process metamodel to improve quantile predictions". In *Proceedings of the 2017 Winter Simulation Conference*, edited by W. K. V. C. et al., 1891–1902. Piscataway, New Jersey: IEEE.

Xie, W., B. L. Nelson, and R. R. Barton. 2014. "A Bayesian framework for quantifying uncertainty in stochastic simulation". *Operations Research* 62(6):1439–1452.

Yi, Y., and W. Xie. 2016. "A simulation-based prediction framework for two-stage dynamic decision making". In *Proceedings of the 2016 Winter Simulation Conference*, edited by T. M. K. R. et al., 2304–2315. Piscataway, New Jersey: IEEE.

Zhou, Q., P. Z. G. Qian, and S. Zhou. 2011. "A simple approach to emulation for computer models with qualitative and quantitative factors". *Technometrics* 53:266–273.

AUTHOR BIOGRAPHIES

WEI XIE is an assistant professor in the Department of Industrial and Systems Engineering at Rensselaer Polytechnic Institute. She received her M.S. and Ph.D. in Industrial Engineering and Management Sciences at Northwestern University. Her research interests are in computer simulation, risk management and data analytics. Her email address is xiew3@rpi.edu and her web page is http://homepages.rpi.edu/~xiew3/.

BO WANG is a Ph.D. candidate of the Department of Industrial and Systems Engineering at Rensselaer Polytechnic Institute. His research interests are input modeling and uncertainty quantification in stochastic simulation. His email address is wangb13@rpi.edu.

QIONG ZHANG is an Assistant Professor of statistics at Virginia Commonwealth University, Richmond, VA. She holds Ph.D. degree in statistics from University of Wisconsin-Madison. Her research interests include computer experiments, uncertainty quantification and spatial and spatial-temporal modeling. She is a member of ASA and INFORMS. Her email address is qzhang4@vcu.edu.