Multivariate Input Uncertainty in Output Analysis for Stochastic Simulation

WEI XIE, Rensselaer Polytechnic Institute
BARRY L. NELSON, Northwestern University
RUSSELL R. BARTON, The Pennsylvania State University

When we use simulations to estimate the performance of stochastic systems, the simulation is often driven by input models estimated from finite real-world data. A complete statistical characterization of system performance estimates requires quantifying both input model and simulation estimation errors. The components of input models in many complex systems could be dependent. In this paper, we represent the distribution of a random vector by its marginal distributions and a dependence measure: either product-moment or Spearman rank correlations. To quantify the impact from dependent input model and simulation estimation errors on system performance estimates, we propose a metamodel-assisted bootstrap framework that is applicable to cases when the parametric family of multivariate input distributions is known or unknown. In either case, we first characterize the input models by their moments that are estimated using real-world data. Then, we employ the bootstrap to quantify the input estimation error, and an equation-based stochastic kriging metamodel to propagate the input uncertainty to the output mean, which can also reduce the influence of simulation estimation error due to output variability. Asymptotic analysis provides theoretical support for our approach, while an empirical study demonstrates that it has good finite-sample performance.

Categories and Subject Descriptors: I.6.6 [**Simulation and Modeling**]: Simulation Output Analysis

General Terms: Algorithms, Experimentation

Additional Key Words and Phrases: Bootstrap, confidence interval, Gaussian process, multivariate input uncertainty, NORTA, output analysis

ACM Reference Format:
Wei Xie, Barry L. Nelson, and Russell R. Barton. 2016. Multivariate input uncertainty in output analysis for stochastic simulation. ACM Trans. Model. Comput. Simul. 27, 1, Article 5 (October 2016), 22 pages.
DOI: http://dx.doi.org/10.1145/2990190

1. INTRODUCTION

Stochastic simulation is used to estimate the performance of complex systems that are driven by random input models. The distributions of these input models are often estimated from finite real-world data. Therefore, a complete statistical characterization of stochastic system performance requires quantifying both input and simulation estimation error. Ignoring either source of uncertainty could lead to unfounded confidence in the system performance estimate. In this paper, we focus on the system mean response,

This is based on work supported by the National Science Foundation under Grant No. CMMI-1068473.
Authors' addresses: W. Xie (corresponding author), Department of Industrial and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY 12180-3590; email: xiew3@rpi.edu. B. L. Nelson, Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60208; email: nelsonb@northwestern.edu. R. R. Barton, Smeal College of Business, Pennsylvania State University, University Park, PA 16802; email: rbarton@psu.edu.

DOI: http://dx.doi.org/10.1145/2990190

and our approach can be applied to other performance estimates, for example, variance and probabilities.

The choice of input models directly impacts the system performance estimates. A prevalent practice is to model the input processes as a collection of independent and identically distributed (i.i.d.) univariate distributions. However, considering that components of real inputs could be dependent, these simple models do not always faithfully represent the physical processes. For example, in a project planning network, the activity durations for different tasks could be correlated if they are affected by the same nuisance factors, for example, weather conditions. In a supply chain system, the demands of a customer for different products, for example, low-fat and whole milk, could be related. In financial risk management, strong dependence between assets in a portfolio could occur if their values are derived from common underlying assets. And in a production-scheduling problem, the operation times for a particular job at a series of processing stations could be dependent. Ignoring such dependence can lead to poor estimates of system performance. Thus, it is desirable to build input models that can faithfully capture the dependence through joint rather than univariate input distributions. In this paper we account for input models with random-vector distributions and do not consider time-series input processes.

Biller and Ghosh [2006] reviewed various approaches to construct joint input distributions. Considering the amount of information needed to specify the joint distribution, almost all these methods suffer from some serious drawbacks. In light of this difficulty, most input-modeling research focuses on methods that match only certain key properties of the input models, including the marginal distributions and some dependence measure.

We assume that dependent input models are characterized by their marginal distributions and a dependence measure. Specifically, the marginal distributions have known parametric families with parameter values unknown. The dependence between different components of input models can be measured by various criteria [Biller and Ghosh 2006]. We focus on product-moment and Spearman rank correlations in this paper. Product-moment correlation measures linear dependence and it is widely used in engineering applications. The definition of product-moment correlation needs the variances of the components to be finite. Thus, we also include the use of Spearman rank correlation as a dependence measure, which finds wide application in business studies, for example, decision and risk analysis [Clemen and Reilly 1999]. Instead of measuring linear dependence, the Spearman rank correlation captures monotonic, possibly nonlinear dependence between different components of input models; it does not require the variances of the components to be finite. Further, since it is based on ranks, it is not sensitive to observation outliers. Notice that in general, dependence may be more than just pairwise and monotonic, in which case a more complex characterization may be needed [Wu and Mielniczuk 2010].

Since marginal distributions and dependence measures are estimated from real-world data, their estimation error is called *input uncertainty*. Therefore, when we use the simulation outputs to estimate system performance, there are two sources of uncertainty: input and simulation error. To quantify the overall uncertainty about the system performance estimate, we build on Xie et al. [2015], which proposed a metamodel-assisted bootstrapping approach to form a confidence interval (CI) accounting for the impact of input and simulation uncertainty. Further, a variance decomposition was proposed to estimate the relative contribution of input to overall uncertainty. However, our previous study in Xie et al. [2015] was based on the assumption that the input distributions are univariate and mutually independent. The independence assumption does not hold in general for input models in many stochastic systems.

This is a significant enhancement of Xie et al. [2015]. To efficiently and correctly account for uncertainty in multivariate input models, we introduce a more general metamodel-assisted bootstrapping framework; it can quantify the impact of dependent input uncertainty and simulation estimation error on system performance estimates while also reducing the influence of simulation estimation error due to finite simulation effort.

Suppose that input models are characterized by marginal distributions and a correlation matrix, and that the parameters are specified by a vector of moments, called moment-based parameters; see Section 4.1 for the detailed definition. In this paper, we consider two cases. First, when the full parametric joint distribution is known except for the marginal distribution parameters and a correlation matrix, then we work with the joint distribution directly, for example, multivariate Pearson distribution. Second, when the parametric joint input distribution is unknown, we construct the joint distribution by using the flexible NORmal To Anything (NORTA) representation [Cario and Nelson 1997]. Since moment-based parameters are estimated with real-world data, the bootstrap is used to quantify the input estimation error, and an equation-based stochastic kriging (SK) metamodel propagates the input uncertainty to the output mean. We can derive a CI that accounts for both simulation and input uncertainty by using this generalized metamodel-assisted bootstrapping approach. Therefore, our approach allows us to do statistical uncertainty analysis for stochastic systems with dependence in the input models.

There are two central contributions of this : first, we generalize the metamodel-assisted bootstrap framework [Xie et al. 2015] to stochastic simulation with dependent input models; second, we propose a rigorous analysis for cases where the dependence is measured by product-moment or Spearman rank correlation.

The next section describes other research on dependent input modeling and input uncertainty analysis. This is followed by a formal description of the problem of interest in Section 3. In Section 4, we propose a generalized metamodel-assisted bootstrapping framework and provide a procedure to build a CI accounting for both input and simulation estimation error on system mean performance estimates. Our approach is supported by asymptotic analysis. We then report results of finite sample behavior from an empirical study in Section 5 and conclude in Section 6. All proofs are in the Online Appendix.

2. BACKGROUND

For stochastic simulations, various approaches to account for input uncertainty have been proposed; see Barton [2012] and Song et al. [2014] for reviews. The methods can be divided into Bayesian and frequentist approaches, which have their underlying merits and limitations [Xie et al. 2014a].

Johnson [1987] reviewed various parametric joint distributions useful in the simulation that can be parameterized by marginal moments and a correlation matrix. For example, the multivariate Johnson translation system matches the first four moments for each marginal distribution and a correlation matrix. A flexible bivariate Gamma distribution proposed in Schmeiser and Lal [1982] allows any Gamma marginal distributions and associated correlations. The multivariate Pearson type II distribution is characterized by the marginal means and a covariance matrix.

When only the marginal distribution families are known, Cario and Nelson [1997] proposed a flexible NORTA distribution to represent and generate random vectors with almost arbitrary marginal distributions and product-moment correlation matrix. Clemen and Reilly [1999] used NORTA to represent the dependent input models for decision and risk analysis with dependence measured by Spearman's and Kendall's rank correlations.

Biller and Corlu [2011] proposed a Bayesian approach to account for the parameter uncertainty for dependent input models. Correlated inputs are modeled with NORTA and the dependence is measured by product-moment correlation. The uncertainty around the NORTA distribution parameters estimated from real-world data is quantified by posterior distributions. For complex stochastic systems with a large number of correlated inputs, a fast algorithm draws samples from these posterior distributions to quantify the input uncertainty. Then, the *direct simulation method* is used to propagate the input uncertainty to the output mean by running simulations at each sample point, which could be computationally expensive for complex simulated systems. Further, the direct simulation method does not incorporate the simulation uncertainty into the Bayesian formulation; see Xie et al. [2014a].

Direct bootstrapping uses bootstrap resampling of the real-world data to represent the input uncertainty and propagates it to the output mean by direct simulation [Barton and Schruben 1993, 2001]; Barton 2007; Cheng and Holland 1997]. Compared with the Bayesian approaches, the direct bootstrap can be adapted to any input models without additional analysis, and it does not need to resort to computationally expensive approaches to draw posterior samples to quantify the input uncertainty. However, direct simulation cannot efficiently use the computational budget to reduce the impact from simulation estimation error. Further, since the statistic that is bootstrapped is the random output of a simulation, it is not a smooth function of input data; this violates the asymptotic validity of the bootstrap.

The metamodel-assisted bootstrapping approach was introduced by Barton et al. [2014]. The input uncertainty is measured by bootstrapping and an equation-based SK metamodel propagates the input uncertainty to the output mean. This approach addresses some of the shortcomings of the direct bootstrap. Specifically, the metamodel can reduce the impact of simulation estimation error. Further, metamodeling makes the bootstrap statistic a smooth function of the input data so that the asymptotic validity concerns faced by the direct bootstrap method disappear. However, Barton et al. [2014] assumed that the simulation budget is not tight and the metamodel uncertainty can be ignored. If the true mean response surface is complex, especially for high-dimensional problems with many input distributions, and the computational budget is tight, then the impact of metamodel uncertainty can no longer be ignored.

The metamodel-assisted bootstrapping approach was improved in Xie et al. [2015] to build a CI accounting for the impact from both input and metamodel uncertainty on the system mean estimates. Further, a variance decomposition was proposed to estimate the relative contribution of input to overall uncertainty, which is very useful for decision makers to determine where to put more effort to reduce the estimator error. The metamodel-assisted bootstrapping approach demonstrates robust performance even when there is a tight computational budget and simulation estimation error is large. However, Xie et al. [2015] is based on the assumption that input models are a collection of mutually independent univariate distributions.

The success of metamodel-assisted bootstrapping for stochastic simulations with independent univariate input distributions in Xie et al. [2015] motivates us to extend it to more complex cases with dependence in the input models.

3. PROBLEM STATEMENT

The stochastic simulation output is a function of random numbers and the input model denoted by F. For notation simplification, we do not explicitly include the random numbers. The output from the jth replication of a simulation with input model F can be written as

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

where $\mu(F) = \mathrm{E}[Y_j(F)]$ 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 model.

In general, F could be composed of mutually independent univariate and multivariate joint distributions. For simplification, suppose that F is composed of a single multivariate distribution with dimension $d > 1$. The marginal distributions of F are denoted by $\{F_1, F_2, \ldots, F_d\}$. In this paper, we focus on continuous marginal distributions.

Suppose that F is characterized by the marginal distributions and a dependence measure: either product-moment or Spearman rank correlation matrix. Specifically, let a $d \times 1$ random vector $\mathbf{X} \sim F$ having $d \times d$ product-moment and Spearman rank correlation matrix denoted, respectively, by $\rho_{\mathbf{X}}$ and $R_{\mathbf{X}}$ with

$$\rho_{\mathbf{X}}(i,j) = \mathrm{corr}(X_i, X_j) = \frac{\mathrm{Cov}(X_i, X_j)}{\sqrt{\mathrm{Var}(X_i)\mathrm{Var}(X_j)}},$$

$$R_{\mathbf{X}}(i,j) = \mathrm{corr}(F_i(X_i), F_j(X_j)) = \frac{\mathrm{E}[F_i(X_i)F_j(X_j)] - \mathrm{E}[F_i(X_i)]\mathrm{E}[F_j(X_j)]}{\sqrt{\mathrm{Var}(F_i(X_i))\mathrm{Var}(F_j(X_j))}}$$

for $i, j = 1, 2, \ldots, d$. Suppose these correlation matrices are positive definite. Since the correlation matrices are symmetric and their diagonal terms are 1, we can view a $d \times d$ correlation matrix as an element of $d^* \equiv d(d-1)/2$ dimensional Euclidean space. Therefore, the product-moment and Spearman rank correlation matrix can be uniquely specified by $d^* \times 1$ vectors denoted by $\mathbf{V}_{\mathbf{X}}^{\rho}$ and $\mathbf{V}_{\mathbf{X}}^{R}$, respectively.

We assume that the families of marginal distributions $\{F_1, F_2, \ldots, F_d\}$ are known, but not their parameter values. Let an $h_i \times 1$ vector $\boldsymbol{\theta}_i$ denote the unknown parameters for the ith marginal distribution F_i. By stacking $\boldsymbol{\theta}_i$ with $i = 1, 2, \ldots, d$ together, we have a $d^\dagger \times 1$ dimensional parameter vector $\boldsymbol{\theta}^\top \equiv (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \ldots, \boldsymbol{\theta}_d^\top)$ with $d^\dagger \equiv \sum_{i=1}^{d} h_i$.

Input models characterized by marginal distributions and correlation matrices can be specified by $\boldsymbol{\vartheta} \equiv (\boldsymbol{\theta}; \mathbf{V}_{\mathbf{X}})$ that includes $d' \equiv d^\dagger + d^*$ elements, where $\mathbf{V}_{\mathbf{X}} = \mathbf{V}_{\mathbf{X}}^{\rho}$ or $\mathbf{V}_{\mathbf{X}}^{R}$. We call $\boldsymbol{\vartheta}$ the *input model parameters*. By abusing notation, we can rewrite $\mu(F)$ as $\mu(\boldsymbol{\vartheta})$. The true input parameters $\boldsymbol{\vartheta}^c$ are unknown and estimated from finite samples of real-world data. Thus, our goal is finding a $(1-\alpha)100\%$ CI $[Q_L, Q_U]$ such that

$$\Pr\{\mu(\boldsymbol{\vartheta}^c) \in [Q_L, Q_U]\} = 1 - \alpha. \tag{1}$$

The unknown input model parameters are estimated by the real-world data, denoted by $\mathbf{X}_m \equiv \{\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(m)}\}$, where the $d \times 1$ random vector $\mathbf{X}^{(i)} \overset{i.i.d}{\sim} F^c$ with $i = 1, 2, \ldots, m$. Under the assumption that the first h_i marginal moments are finite for $i = 1, 2, \ldots, d$, we estimate marginal distribution parameters by the moment estimators, denoted by $\widehat{\boldsymbol{\theta}}_m$ [Xie et al. 2015]. The usual estimators for the product-moment and Spearman rank correlations are

$$\widehat{\rho}_{\mathbf{X},m}(i,j) = \frac{\sum_{k=1}^{m}\left(X_i^{(k)} - \bar{X}_i\right)\left(X_j^{(k)} - \bar{X}_j\right)/(m-1)}{S_i S_j} \tag{2}$$

$$\widehat{R}_{\mathbf{X},m}(i,j) = \frac{\sum_{k=1}^{m}\left(\mathrm{r}(X_i^{(k)}) - \overline{\mathrm{r}(X_i)}\right)\left(\mathrm{r}(X_j^{(k)}) - \overline{\mathrm{r}(X_j)}\right)}{\sqrt{\left[\sum_{k=1}^{m}\left(\mathrm{r}(X_i^{(k)}) - \overline{\mathrm{r}(X_i)}\right)^2\right]\cdot\left[\sum_{k=1}^{m}\left(\mathrm{r}(X_j^{(k)}) - \overline{\mathrm{r}(X_j)}\right)^2\right]}} \tag{3}$$

for $i, j = 1, 2, \ldots, d$, where $\bar{X}_i = \sum_{k=1}^{m} X_i^{(k)}/m$ and $S_i^2 = \sum_{k=1}^{m}\left(X_i^{(k)} - \bar{X}_i\right)^2/(m-1)$. We denote the rank function by $\mathrm{r}(\cdot) \equiv \mathrm{rank}(\cdot)$. In this paper, we use the uprank to

estimate the Spearman rank correlations. It is defined by $\mathrm{r}(X_i^{(k)}) \equiv \sum_{j=1}^{m} \mathrm{I}(X_i^{(j)} \leq X_i^{(k)})$ and $\overline{\mathrm{r}(X_i)} = \sum_{k=1}^{m} \mathrm{r}(X_i^{(k)})/m$ with $\mathrm{I}(\cdot)$ denoting an indicator function. Then, based on Equations (2) and (3), we can find corresponding correlation estimators $\widehat{\mathbf{V}}^{\rho}_{\mathbf{X},m}$ and $\widehat{\mathbf{V}}^{R}_{\mathbf{X},m}$. Given the input-model parameter estimator $\widehat{\boldsymbol{\vartheta}}_m = (\widehat{\boldsymbol{\theta}}_m; \widehat{\mathbf{V}}^{\rho}_{\mathbf{X},m})$ or $(\widehat{\boldsymbol{\theta}}_m; \widehat{\mathbf{V}}^{R}_{\mathbf{X},m})$ that is a function of real-world data $\mathbf{X}_m$, its sampling distribution can be used to quantify input uncertainty.

The impact of input uncertainty on the system mean performance estimate is quantified by the sampling distribution of $\mu(\widehat{\boldsymbol{\vartheta}}_m)$. Further, since the underlying response surface $\mu(\cdot)$ is unknown, at any $\widehat{\boldsymbol{\vartheta}}$, let $\widehat{\mu}(\widehat{\boldsymbol{\vartheta}})$ denote the corresponding mean response estimator. Thus, there are both input and simulation estimation errors in the system mean performance estimates.

For stochastic systems with dependent input models, our objective is to create an approach to quantify the overall impact of both input and simulation estimation error on system mean performance estimates and then build a CI satisfying Equation (1). Further, since each simulation run could be computationally expensive and we may have a tight computational budget, we want to reduce the influence of simulation estimation error.

4. METAMODEL-ASSISTED BOOTSTRAPPING FRAMEWORK

For problems with parametric input distributions that are univariate and mutually independent, the metamodel-assisted bootstrapping framework was used to account for the impact of both input and simulation estimation errors on the system performance estimates in Xie et al. [2015]. In this section, we generalize the metamodel-assisted bootstrapping approach for stochastic simulations with dependence in the input models.

To make this section easy to follow, we start with an overall description of the generalized metamodel-assisted bootstrapping framework. We employ the bootstrap to capture the estimation error of moment-based input parameters in Section 4.1, and propagate the input uncertainty to the output mean by using an equation-based stochastic kriging metamodel that is built based on the simulation outputs at a few well-chosen design points; see Section 4.3. At each design point, we need to construct a full joint input distribution, generate samples of $\mathbf{X}$ to drive simulations, and estimate the system mean responses. We consider cases with the family of parametric joint distribution known or unknown, respectively, in Section 4.2. Then, since both simulation and metamodel uncertainty can be estimated using properties of an SK metamodel, we propose a procedure to deliver a CI that accounts for both simulation and input uncertainty in Section 4.4. Asymptotic analysis provides theoretical support for our approach.

Notice that direct simulation, by running the simulations at each bootstrapped sample of input moments to estimate the system performance, could be used to propagate the input uncertainty to the output. Our previous study [Xie et al. 2014a] demonstrates the advantages by using an SK metamodel over direct simulation. Given a finite simulation budget, the SK metamodel efficiently reduces the impact of simulation estimation uncertainty. In addition, bootstrap consistency requires the statistic to be a smooth function of the data. Direct simulation violates this requirement. This issue does not exist in our metamodel-assisted bootstrapping approach; see Barton et al. [2014] and Xie et al. [2015].

4.1. Bootstrap for Input Uncertainty

In this section, we describe how to employ the bootstrap to quantify the input uncertainty. The way we choose to represent input models plays an important role in

the implementation of metamodel-assisted bootstrapping. *Moment-based parameters*, denoted by $\mathcal{M}$, are used to characterize the input model with dependence. Instead of using the natural parameters $\boldsymbol{\theta}$ to characterize the marginal distributions, we can use moments; see Barton et al. [2014] for an explanation. Suppose that the parametric marginal distribution F_i can be uniquely characterized by its first h_i finite moments denoted by the $h_i \times 1$ vector $\boldsymbol{\psi}_i$ for $i = 1, 2, \ldots, d$. By stacking $\boldsymbol{\psi}_i$ with $i = 1, 2, \ldots, d$ together, we have a $d^\dagger \times 1$ dimensional vector of marginal moments $\boldsymbol{\psi}^\top \equiv (\boldsymbol{\psi}_1^\top, \boldsymbol{\psi}_2^\top, \ldots, \boldsymbol{\psi}_d^\top)$. Therefore, the input models can be characterized by the collection of moments $\mathcal{M} = (\boldsymbol{\psi}; \mathbf{V}_\mathbf{X})$ with $\mathbf{V}_\mathbf{X} = \mathbf{V}_\mathbf{X}^\rho$ or $\mathbf{V}_\mathbf{X}^R$. Suppose there is a one-to-one continuous mapping between marginal moments and parameters, denoted by $\boldsymbol{\theta} = h(\boldsymbol{\psi})$. Thus, the input parameters $\boldsymbol{\vartheta}$ and moments $\mathcal{M}$ are interchangeable. Abusing notation again, we rewrite $\mu(\boldsymbol{\vartheta})$ as $\mu(\mathcal{M})$.

The true moments of dependent input models, denoted by $\mathcal{M}^c$, are unknown and estimated based on a finite sample $\mathbf{X}_m$. Specifically, we use standardized sample moments as estimators for marginal distributions, denoted by $\widehat{\boldsymbol{\psi}}_m$; see Xie et al. [2015]. The correlation estimator $\widehat{\mathbf{V}}_{\mathbf{X},m}^\rho$ or $\widehat{\mathbf{V}}_{\mathbf{X},m}^R$ is obtained by using Equation (2) or (3). The estimation error of input models can be quantified by the sampling distribution of $\widehat{\mathcal{M}}_m = (\widehat{\boldsymbol{\psi}}_m; \widehat{\mathbf{V}}_{\mathbf{X},m}^\rho)$ or $(\widehat{\boldsymbol{\psi}}_m; \widehat{\mathbf{V}}_{\mathbf{X},m}^R)$, denoted by $F_{\mathcal{M}_m}^c$. Therefore, the impact of input uncertainty on the system mean performance estimate can be measured by the sampling distribution of $\mu(\widehat{\mathcal{M}}_m)$ with $\widehat{\mathcal{M}}_m \sim F_{\mathcal{M}_m}^c$.

Since it could be hard to derive the sampling distribution $F_{\mathcal{M}_m}^c$, we use bootstrap resampling to approximate it [Shao and Tu 1995]. Let $A \equiv \{1, 2, \ldots, m\}$. Implementation of bootstrap resampling in the metamodel-assisted bootstrapping is as follows:

(1) Draw m samples with replacement from set A and obtain bootstrapped indexes $\{i_1, i_2, \ldots, i_m\}$; choose corresponding samples from real-world data $\mathbf{X}_m$ and get $\widetilde{\mathbf{X}}_m^{(1)} \equiv \{\mathbf{X}^{(i_1)}, \mathbf{X}^{(i_2)}, \ldots, \mathbf{X}^{(i_m)}\}$. Use $\widetilde{\mathbf{X}}_m^{(1)}$ to calculate the bootstrapped moment estimate, denoted by $\widetilde{\mathcal{M}}_m^{(1)} \equiv (\widetilde{\boldsymbol{\psi}}_m^{(1)}; (\widetilde{\mathbf{V}}_{\mathbf{X},m}^\rho)^{(1)})$ or $(\widetilde{\boldsymbol{\psi}}_m^{(1)}; (\widetilde{\mathbf{V}}_{\mathbf{X},m}^R)^{(1)})$.
(2) Repeat Step (1) for B times to generate $\widetilde{\mathcal{M}}_m^{(b)}$ with $b = 1, 2, \ldots, B$.

The bootstrap resampled moments are drawn from the bootstrap distribution, denoted by $\widetilde{F}_{\mathcal{M}_m}(\cdot|\mathbf{X}_m)$, with $\widetilde{\mathcal{M}}_m \sim \widetilde{F}_{\mathcal{M}_m}(\cdot|\mathbf{X}_m)$. For estimation of a percentile CI quantifying the impact of input uncertainty, B is recommended to be a few thousand; see Xie et al. [2014a]. In this paper, a $\widehat{\ }$ denotes a quantity estimated from real-world data, while a $\widetilde{\ }$ denotes a quantity estimated from bootstrapped data.

Theorem 4.1 shows that when the amount of real-world data increases to infinity, the bootstrap provides a consistent estimator for the true input moments $\mathcal{M}^c$.

THEOREM 4.1. *Suppose the following conditions hold:*

(*1*) *We have* $\mathbf{X}^{(k)} \overset{i.i.d}{\sim} F^c$ *with* $k = 1, 2, \ldots, m$.
(*2*) *The marginal distribution* F_i^c *is uniquely characterized by its first* h_i *moments and it has finite first* $4h_i$ *moments for* $i = 1, 2, \ldots, d$.
(*3*) $\mathrm{E}(X_i^4 X_j^4) < \infty$ *for* $i, j = 1, 2, \ldots, d$.

Then, as $m \to \infty$, *the bootstrap moment estimator* $\widetilde{\mathcal{M}}_m$ *converges a.s. to the true moments* $\mathcal{M}^c$.

The proof of Theorem 4.1 is provided in the Online Appendix.

Notice that under some situations, such as when the input model does not have enough finite moments, the normal approximation obtained by the Central Limit Theorem may perform better than the bootstrap; see Hall [1988]. For this case, we could

easily extend our framework by using the normal approximation to quantify the input uncertainty.

4.2. Construction of Joint Input Distributions

Given feasible moment-based parameters, we describe the procedure to construct the joint input distribution in this section. We first consider the case when the full parametric joint distribution is known except the marginal parameters and a correlation matrix in Section 4.2.1. Then, when the parametric family is unknown, we use a NORTA representation to construct the joint distribution in Section 4.2.2. Notice that unless the true distribution is NORTA, there is unmeasured error due to incorrect input models. That error is not addressed in this paper.

4.2.1. Parametric Joint Input Distributions. In this section, we consider multivariate parametric input distributions F with the distribution family known. The input model is specified by marginal parameters $\boldsymbol{\theta}$ and a correlation matrix $\rho_{\mathbf{X}}$; see Schmeiser and Lal [1982] and Johnson [1987] for multivariate families useful in simulation. The underlying correct parameters $(\boldsymbol{\theta}^c; (\mathbf{V}_{\mathbf{X}}^{\rho})^c)$ are unknown and estimated by finite real-world data.

We use a multivariate Pearson type II distribution [Johnson 1987] as an illustrative example. It has the density function

$$f(\mathbf{x}) = \frac{\Gamma(d/2+\kappa+1)}{\Gamma(\kappa+1)\pi^{d/2}}|\Sigma|^{-1/2}[1-(\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})]^{\kappa}.$$

Suppose that the shape parameter κ is given. Then, the Pearson type II distribution is specified by parameters $\boldsymbol{\mu}$ and Σ with

$$\mathrm{E}(\mathbf{X}) = \boldsymbol{\mu} \text{ and } \mathrm{Cov}(\mathbf{X}) = \frac{\Sigma}{2\kappa+d+2}.$$

It can also be specified by the moment vector $\mathcal{M}$, including the first two marginal moments and the correlation vector $\mathbf{V}_{\mathbf{X}}^{\rho}$.

For a general parametric joint distribution F that could be specified by marginal parameters $\boldsymbol{\theta}$ and a correlation matrix $\rho_{\mathbf{X}}$, Theorem 4.2 shows that for any moment vector $\mathcal{M}$ in a small neighborhood centered at $\mathcal{M}^c$, we could find a feasible multivariate parametric distribution F.

THEOREM 4.2. *Let F be a parametric multivariate input distribution specified by marginal parameters $\boldsymbol{\theta}$ and a correlation matrix $\rho_{\mathbf{X}}$. Let $\boldsymbol{\Theta} \subseteq \Re^{d'}$ be the feasible domain for marginal distribution parameters $\boldsymbol{\theta}$. Suppose the following conditions hold:*

(1) *Any parameter vector $\boldsymbol{\vartheta}$ with $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and positive semidefinite $\rho_{\mathbf{X}}$ has a feasible multivariate parametric joint distribution F.*
(2) *$\boldsymbol{\theta}^c$ is an interior point in $\boldsymbol{\Theta}$, and $\rho_{\mathbf{X}}^c$ is positive definite.*
(3) *There is a one-to-one continuous mapping between marginal moments and parameters, $\boldsymbol{\theta} = h(\boldsymbol{\psi})$.*

Then the true moment vector $\mathcal{M}^c = (\boldsymbol{\psi}^c; (\mathbf{V}_{\mathbf{X}}^{\rho})^c)$ is an interior point of the feasible region: in the d' dimensional Euclidean space, there exists a constant $\delta > 0$ such that any moment combination $\mathcal{M}$ in the open ball $B_{\delta}(\mathcal{M}^c)$ has a feasible parametric multivariate distribution F.

The proof of Theorem 4.2 is provided in the Online Appendix.

Notice that if the marginals have the same family, we may find an existing parametric multivariate distribution to use; see Johnson [1987]. For marginals having different families, we typically cannot use a standard multivariate parametric distribution, and

we need to consider the transformation-based approaches, for example, NORTA and multivariate Johnson distributions.

4.2.2. NORTA Representation. In this section, suppose that the parametric family of joint input distribution F is unknown. Given the partial characterization specified by marginal distributions and a pairwise dependence measure, either product-moment or Spearman rank correlations, we now describe how to employ the NORTA representation for constructing the joint distributions and generating samples of random vector $\mathbf{X}$.

To find a NORTA representation for F, we represent $\mathbf{X}$ as a transformation of a d-dimensional standard multivariate normal (MVN) vector $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_d)^\top$ with product-moment correlation matrix denoted by $\rho_{\mathbf{Z}}$,

$$\mathbf{X}^\top = \left(F_1^{-1}[\Phi(Z_1); \boldsymbol{\theta}_1], F_2^{-1}[\Phi(Z_2); \boldsymbol{\theta}_2], \ldots, F_d^{-1}[\Phi(Z_d); \boldsymbol{\theta}_d]\right), \tag{4}$$

where $\Phi(\cdot)$ denotes the cdf for the standard normal distribution. If the marginal distribution families are given, as we assume here, then the NORTA representation for F can be specified by $(\boldsymbol{\theta}, \rho_{\mathbf{Z}})$. For a standard normal random vector $\mathbf{Z}$, there is a closed-form relation between product-moment correlation $\rho_{\mathbf{Z}}$ and Spearman rank correlation $R_{\mathbf{Z}}$ [Clemen and Reilly 1999]:

$$R_{\mathbf{Z}}(i, j) = \frac{6}{\pi} \sin^{-1}\left(\frac{\rho_{\mathbf{Z}}(i, j)}{2}\right), \tag{5}$$

with $i, j = 1, 2, \ldots, d$.

Since the NORTA implementations for cases where the dependence is measured by Spearman rank and product-moment correlations are different, we describe them separately.

If the dependence in the input models is measured by the Spearman rank correlation, we have $R_{\mathbf{X}} = R_{\mathbf{Z}}$ since it is invariant under monotone one-to-one transformation $F_i^{-1}[\Phi(\cdot)]$ for $i = 1, 2, \ldots, d$. Therefore, given moment-based parameters $\mathcal{M} = (\boldsymbol{\psi}; \mathbf{V}_{\mathbf{X}}^R)$, we can find $\boldsymbol{\vartheta} = (\boldsymbol{\theta}; \mathbf{V}_{\mathbf{X}}^R)$ by moment matching. The procedure to find a NORTA representation and generate samples for $\mathbf{X}$ is as follows:

(1) From $\mathbf{V}_{\mathbf{X}}^R$, get the Spearman rank correlation matrix for $\mathbf{Z}$, $R_{\mathbf{Z}} = R_{\mathbf{X}}$, and obtain corresponding product-moment correlation $\rho_{\mathbf{Z}}(i, j) = 2\sin(\pi R_{\mathbf{Z}}(i, j)/6)$ for $i, j = 1, 2, \ldots, d$.

(2) Generate $\mathbf{Z} \overset{i.i.d.}{\sim} \text{MVN}(\mathbf{0}, \rho_{\mathbf{Z}})$ and obtain $\mathbf{X}$ by using Equation (4).

By repeating this procedure, we generate samples for $\mathbf{X}$, use them to drive simulations, and estimate the mean response $\mu(\mathcal{M})$.

Notice that when we use Spearman rank correlation to measure the pairwise dependence between the components of input models, the choice of marginal distributions and correlation is separable. However, for F specified by a combination of feasible $\boldsymbol{\theta}$ and a positive definite Spearman rank correlation matrix $R_{\mathbf{X}}$, we may not find a NORTA representation because the nonlinear transformation of positive definite $R_{\mathbf{Z}}$, $\rho_{\mathbf{Z}}(i, j) = 2\sin[\pi R_{\mathbf{Z}}(i, j)/6]$, could lead to a nonpositive definite correlation matrix $\rho_{\mathbf{Z}}$; see Ghosh and Henderson [2002a] and Li and Hammond [1975].

If the dependence between the components of input models is measured by product-moment correlation, then the procedure to find a NORTA representation becomes more complex because the choice of marginal distributions influences the feasibility of the correlation matrix. Specifically, there is a pairwise relation between product-moment

correlation matrices of $\mathbf{X}$ and $\mathbf{Z}$:

$$\begin{aligned}\rho_{\mathbf{X}}(i,j) &= C_{ij}[\rho_{\mathbf{Z}}(i,j);\boldsymbol{\theta}] \\ &\equiv \frac{\int\int F_i^{-1}[\Phi(z_i);\boldsymbol{\theta}_i]F_j^{-1}[\Phi(z_j);\boldsymbol{\theta}_j]\varphi_{\rho_{\mathbf{Z}}(i,j)}(z_i,z_j)\mathrm{d}z_i\mathrm{d}z_j - \mathrm{E}(X_i)\mathrm{E}(X_j)}{\sqrt{\mathrm{Var}(X_i)\mathrm{Var}(X_j)}},\end{aligned} \tag{6}$$

where $\varphi_{\rho_{\mathbf{Z}}(i,j)}$ denotes the standard bivariate normal density with correlation $\rho_{\mathbf{Z}}(i,j)$ and C_{ij} denotes a pairwise transformation from $\rho_{\mathbf{Z}}(i,j)$ to $\rho_{\mathbf{X}}(i,j)$ for $i,j = 1,2,\ldots,d$. Given $\boldsymbol{\theta}$ and $\rho_{\mathbf{X}}$, we solve d^* correlation-matching Equation (6) to find $\rho_{\mathbf{Z}}$. Unlike Spearman rank correlation, the marginal distributions characterized by parameters $\boldsymbol{\theta}$ play an important role in determining the value and feasibility of correlation matrix $\rho_{\mathbf{Z}}$.

Given $\mathcal{M} = (\boldsymbol{\psi};\mathbf{V}_{\mathbf{X}}^{\rho})$, we can find $\boldsymbol{\vartheta} = (\boldsymbol{\theta};\mathbf{V}_{\mathbf{X}}^{\rho})$ by moment matching. If there exists a feasible NORTA representation, the procedure to find it and generate samples for $\mathbf{X}$ is as follows:

(1) Given $\mathbf{V}_{\mathbf{X}}^{\rho}$, solve Equation (6) for correlation matrix $\rho_{\mathbf{Z}}$.

(2) Generate $\mathbf{Z} \overset{i.i.d.}{\sim} \mathrm{MVN}(\mathbf{0}, \rho_{\mathbf{Z}})$ and obtain $\mathbf{X}$ by using Equation (4).

By repeating this procedure, we generate samples for $\mathbf{X}$ and then use them to drive simulations to estimate $\mu(\mathcal{M})$. When we solve for $\rho_{\mathbf{Z}}$ in Step (1), typically there is no closed-form analytical solution except for some special marginal distributions, for example, uniform distribution, and we need to resort to numerical search to obtain $\rho_{\mathbf{Z}}$.

Notice that when we use product-moment correlation to characterize the pairwise dependence between the components of input models, the nonlinear transformation between $\rho_{\mathbf{X}}$ and $\rho_{\mathbf{Z}}$ in Equation (6) may not guarantee the positive semidefinite property for $\rho_{\mathbf{Z}}$. There *may not exist* a NORTA representation for every feasible combination of $\boldsymbol{\theta}$ and positive definite $\rho_{\mathbf{X}}$. Therefore, the previous procedure only works under the condition that the correlation matrix $\rho_{\mathbf{Z}}$ obtained in Step (1) is positive semidefinite, which may not hold in general. Based on the study by Ghosh and Henderson [2002b], this infeasibility is more likely in high dimensions and with correlations close to ± 1. When this happens, we can find a NORTA feasible correlation matrix that is close to $\rho_{\mathbf{X}}$.

If $\rho_{\mathbf{Z}}$ obtained by nonlinear transformation $\rho_{\mathbf{Z}}(i,j) = 2\sin[\pi R_{\mathbf{Z}}(i,j)/6]$ or by solving d^* in Equation (6) is positive semidefinite, $\boldsymbol{\vartheta} = (\boldsymbol{\theta};\mathbf{V}_{\mathbf{X}})$ with $\mathbf{V}_{\mathbf{X}} = \mathbf{V}_{\mathbf{X}}^{R}$ or $\mathbf{V}_{\mathbf{X}}^{\rho}$ is called *NORTA feasible*; otherwise, it is called *NORTA infeasible*. Theorem 4.3 gives a property of the NORTA feasible region: *if the true moment combination $\mathcal{M}^c$ has a NORTA feasible representation with positive definite $\rho_{\mathbf{Z}}^c$, then there exists a neighborhood centered at $\mathcal{M}^c$ in the $d^\dagger + d^*$ dimensional Euclidean space such that every moment combination $\mathcal{M}$ in the neighborhood has a NORTA feasible representation.* This property is useful when we show the asymptotic consistency of the CI built by the metamodel-assisted bootstrapping to quantify both input and simulation uncertainty in Section 4.4.

THEOREM 4.3. *Let $\boldsymbol{\Theta} \subseteq \Re^{d^\dagger}$ be the feasible domain for marginal distribution parameters $\boldsymbol{\theta}$ and suppose $\boldsymbol{\theta}^c$ is an interior point in $\boldsymbol{\Theta}$. Suppose there is one-to-one continuous mapping between marginal moments $\boldsymbol{\psi}_i$ and parameters $\boldsymbol{\theta}_i$ for $i = 1,2,\ldots,d$. Suppose the following conditions hold:*

(1) F^c has a NORTA feasible representation $(\boldsymbol{\theta}^c, \rho_{\mathbf{Z}}^c)$ with $\rho_{\mathbf{Z}}^c$ positive definite.
(2) At any x, the marginal distributions $F_i(x;\boldsymbol{\theta}_i)$, density functions $f_i(x;\boldsymbol{\theta}_i)$, and inverse distributions $F_i^{-1}(x;\boldsymbol{\theta}_i)$ are continuously differentiable over $\boldsymbol{\theta}_i$ for $i = 1,2,\ldots,d$ on $\boldsymbol{\Theta}$.

(3) *For any* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$*, the marginal cdfs* $F_1(x;\boldsymbol{\theta}_1), F_2(x;\boldsymbol{\theta}_2), \ldots, F_d(x;\boldsymbol{\theta}_d)$ *are continuous and strictly increasing in* x*.*

Then the true moment vector $\mathcal{M}^c = (\boldsymbol{\psi}^c; \mathbf{V}_{\mathbf{X}}^c)$ *with* $\mathbf{V}_{\mathbf{X}} = \mathbf{V}_{\mathbf{X}}^\rho$ *or* $\mathbf{V}_{\mathbf{X}}^R$ *is an interior point of the NORTA feasible region: in the* $d^\dagger + d^*$ *dimensional Euclidean space, there exists a constant* $\delta > 0$ *such that every moment combination* $\mathcal{M}$ *in the open ball* $B_\delta(\mathcal{M}^c)$ *has a NORTA feasible representation.*

The proof of Theorem 4.3 is provided in the Online Appendix.

4.3. Stochastic Kriging Metamodel

In the metamodel-assisted bootstrapping framework, after quantifying the input uncertainty with the bootstrap as described in Section 4.1, an equation-based SK metamodel introduced by Ankenman et al. [2010] is used to propagate the input uncertainty to the output mean. The succinct review of SK in this section is based on Ankenman et al. [2010].

Dependent input models characterized by finite moment-based parameters $\mathcal{M}$ composed of marginal standard moments and component-pairwise correlations can be interpreted as a location $\mathbf{x}$ in a $d' = (d^\dagger + d^*)$ dimension space. The p-norm distance could be used to measure the difference between moment-based parameters $\mathbf{x}$ and $\mathbf{x}'$ defined by

$$d(\mathbf{x}, \mathbf{x}') = \|\mathbf{x}^\top \boldsymbol{\zeta} \mathbf{x}'\|_p^{1/p} = \left[\sum_{j=1}^{d'} \zeta_j \left(x_j - x_j'\right)^p\right]^{1/p}, \tag{7}$$

with $p \geq 1$, where $\mathbf{x}$ is a $d' \times 1$ moment vector and $\boldsymbol{\zeta}$ denotes a $d' \times d'$ diagonal matrix with nonnegative diagonal terms $\zeta_1, \zeta_2, \ldots, \zeta_{d'}$. In this paper, the distance between different estimates of input models is measured by a weighted Euclidean distance on moment-based parameters with $p = 2$. Since the marginal moments and pairwise correlations could have different impacts on the system performance, the weights $\boldsymbol{\zeta}$ quantify their relative effects. For example, if the marginal moments have dominant effects, the corresponding weights tend to be higher.

This distance measure was used to measure the difference between moment estimates for univariate parametric input models in Barton et al. [2014]. In terms of distance measures for a correlation matrix, Higham [2002] used weighted Euclidean distance, while Ghosh and Henderson [2002b] used other norms, including L_1 and L_∞. We choose the weighted Euclidean distance because each individual correlation can matter and different correlation elements could have different impacts on the system mean performance. The distance measures L_1 and L_∞ do not capture that. Notice that since the correlation matrix is required to be positive semidefinite, there exist implicit constraints on the elements of the correlation matrix.

The input models with closer moments tend to have closer mean responses. Suppose that the underlying true (but unknown) response surface is a continuous function of moment-based parameters $\mathbf{x}$ and $\mu(\cdot)$ is a realization of a stationary Gaussian Process (GP). We model the simulation output Y by

$$Y_j(\mathbf{x}) = \beta_0 + W(\mathbf{x}) + \epsilon_j(\mathbf{x}). \tag{8}$$

This model includes two sources of uncertainty: the simulation output uncertainty $\epsilon_j(\mathbf{x})$ and mean response uncertainty characterized by the GP $W(\mathbf{x})$. For many, but not all, simulation settings, the output is an average of a large number of more basic outputs, so a normal approximation can be applied: $\epsilon(\mathbf{x}) \sim \mathrm{N}(0, \sigma_\epsilon^2(\mathbf{x}))$. For example, when we study the steady-state expected waiting time in a queue, each simulation output is the average of waiting times for many customers.

Since stochastic systems with dependent input models having similar key properties tend to have close mean responses, a zero-mean, second-order stationary GP $W(\cdot)$ is used to account for this spatial dependence. Therefore, the uncertainty about the unknown true response surface $\mu(\mathbf{x})$ is represented by a GP $M(\mathbf{x}) \equiv \beta_0 + W(\mathbf{x})$ (note that β_0 can be replaced by a more general trend term $\mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta}$). Its spatial dependence is characterized by the covariance function, $\Sigma(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}[W(\mathbf{x}), W(\mathbf{x}')] = \tau^2 \gamma(\mathbf{x} - \mathbf{x}')$, where τ^2 denotes the variance and $\gamma(\cdot)$ is a correlation function that depends only on the distance $\mathbf{x} - \mathbf{x}'$. Based on prior information about the smoothness of $\mu(\cdot)$, we can choose the form of correlation function [Xie et al. 2010]. Considering that mean response surfaces for most system engineering problems have a high order of smoothness and Gaussian correlation function demonstrates good performance [Mukhopadhyay et al. 2016], we use the product-form Gaussian correlation function

$$\gamma(\mathbf{x} - \mathbf{x}') = \exp\left(- \sum_{j=1}^{d'} \zeta_j (x_j - x_j')^2 \right) \tag{9}$$

for the empirical evaluation in Section 5. In SK, the weights $\boldsymbol{\zeta} = (\zeta_1, \zeta_2, \ldots, \zeta_{d'})$ are also called correlation parameters that quantify the relative effects of elements in the moment-based parameters on the system mean response. Before having any simulation result, the uncertainty about $\mu(\mathbf{x})$ can be represented by a Gaussian process $M(\mathbf{x}) \sim \mathrm{GP}(\beta_0, \tau^2 \gamma(\mathbf{x} - \mathbf{x}'))$.

To reduce the uncertainty about $\mu(\mathbf{x})$, we choose an experiment design consisting of pairs $\mathcal{D} \equiv \{(\mathbf{x}_i, n_i), i = 1, 2, \ldots, K\}$ with $(\mathbf{x}_i, n_i)$ denoting the location and the number of replications at the ith design point. The simulation outputs at $\mathcal{D}$ are $\mathbf{Y}_\mathcal{D} \equiv \{(Y_1(\mathbf{x}_i), Y_2(\mathbf{x}_i), \ldots, Y_{n_i}(\mathbf{x}_i)); i = 1, 2, \ldots, K\}$ and the sample mean at design point $\mathbf{x}_i$ is $\bar{Y}(\mathbf{x}_i) = \sum_{j=1}^{n_i} Y_j(\mathbf{x}_i)/n_i$. Let the sample means at all K design points be $\bar{\mathbf{Y}}_\mathcal{D} = (\bar{Y}(\mathbf{x}_1), \bar{Y}(\mathbf{x}_2), \ldots, \bar{Y}(\mathbf{x}_K))^T$ and its variance be represented by a $K \times K$ diagonal matrix $C = \mathrm{diag}\{\sigma_\epsilon^2(\mathbf{x}_1)/n_1, \sigma_\epsilon^2(\mathbf{x}_2)/n_2, \ldots, \sigma_\epsilon^2(\mathbf{x}_K)/n_K\}$ because the use of common random numbers is detrimental to prediction [Chen et al. 2012].

The simulation outputs $\mathbf{Y}_\mathcal{D}$ and spatial dependence characterized by the covariance function $\Sigma(\cdot, \cdot)$ can be used to improve system mean prediction at any fixed point $\mathbf{x}$. Specifically, let Σ be the $K \times K$ spatial covariance matrix of the design points and let $\Sigma(\mathbf{x}, \cdot)$ be the $K \times 1$ spatial covariance vector between each design point and $\mathbf{x}$. If the parameters $(\tau^2, \boldsymbol{\zeta}, C)$ are known, then the metamodel uncertainty can be characterized by a refined GP $M_p(\mathbf{x})$ that denotes the conditional distribution of $M(\mathbf{x})$ given all simulation outputs,

$$M_p(\mathbf{x}) \sim \mathrm{GP}(m_p(\mathbf{x}), \sigma_p^2(\mathbf{x})), \tag{10}$$

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

$$m_p(\mathbf{x}) = \widehat{\beta}_0 + \Sigma(\mathbf{x}, \cdot)^\top (\Sigma + C)^{-1} (\bar{\mathbf{Y}}_\mathcal{D} - \widehat{\beta}_0 \cdot 1_{K \times 1}), \tag{11}$$

and the corresponding variance is

$$\sigma_p^2(\mathbf{x}) = \tau^2 - \Sigma(\mathbf{x}, \cdot)^\top (\Sigma + C)^{-1} \Sigma(\mathbf{x}, \cdot) + \eta^\top [1_{K \times 1}^\top (\Sigma + C)^{-1} 1_{K \times 1}]^{-1} \eta, \tag{12}$$

where $\widehat{\beta}_0 = [1_{K \times 1}^\top (\Sigma + C)^{-1} 1_{K \times 1}]^{-1} 1_{K \times 1}^\top (\Sigma + C)^{-1} \bar{\mathbf{Y}}_\mathcal{D}$ and $\eta = 1 - 1_{K \times 1}^\top (\Sigma + C)^{-1} \Sigma(\mathbf{x}, \cdot)$ [Ankenman et al. 2010]. Notice that misspecified correlation functions could cause biased $\widehat{\beta}_0$.

Since in reality the spatial correlation parameters τ^2 and $\boldsymbol{\zeta}$ are unknown, maximum likelihood estimates (MLEs) are typically used for prediction, and the sample variance is used as an estimate for the simulation variance at design points C [Ankenman et al.

2010]. By substituting parameter estimates $(\widehat{\tau}^2, \widehat{\boldsymbol{\zeta}}, \widehat{C})$ in Equations (11) and (12), we can obtain the estimated mean $\widehat{m}_p(\mathbf{x})$ and variance $\widehat{\sigma}_p^2(\mathbf{x})$. Thus, the metamodel we use is $\widehat{\mu}(\mathbf{x}) = \widehat{m}_p(\mathbf{x})$ with variance estimated by $\widehat{\sigma}_p^2(\mathbf{x})$.

In our study, we do not account for the estimation error of SK parameters $(\tau^2, \boldsymbol{\zeta}, C)$. This is common in the kriging literature because fully including the effect of these parameters' estimation error would make the distribution of SK metamodel $M(\cdot)$ mathematically and computationally intractable. The impact of SK parameter estimation uncertainty on the metamodel fit in a general situation has not been comprehensively studied. The studies in Das et al. [2012] and Bachoc [2013] show the asymptotic consistency of kriging parameter estimates via weighted least square (WLS), MLE, or cross-validation. The studies in Xie et al. [2015] and Yin et al. [2009] indicate that when we use the space-filling design with a reasonable number of design points [Jones et al. 1998; Loeppky et al. 2009] and the simulation estimation uncertainty does not dominate the information from the underlying response surface, the metamodel fit is robust to the SK parameter estimation uncertainty. This does not hold when the model is used for extrapolation. In our study, we construct a design space that covers the most likely bootstrapped input moments, which avoids the extrapolation issue; see Section 4.4.

4.4. Procedure to Build a CI

Since there are both input and simulation estimation errors in the system mean performance estimates, in this section, we propose a procedure to build a CI quantifying the overall uncertainty for $\mu(\mathcal{M}^c)$. We show that as $m, B \to \infty$, the CI has asymptotically consistent coverage.

Based on a hierarchical sampling approach, we propose the following procedure to build a $(1-\alpha)100\%$ bootstrap percentile CI. We do bootstrapping over moment-based parameters to quantify the input uncertainty. Since each simulation run could be expensive, to efficiently propagate the input uncertainty quantified by B bootstrapped moment samples, $\widetilde{\mathcal{M}}_m^{(b)}$ with $b = 1, 2, \ldots, B$, to outputs, we construct an SK metamodel in Steps (1) through (3) that covers the most likely bootstrapped samples. Then, Step (4) uses the SK metamodel to propagate the input uncertainty to output means, with part (a) accounting for the input uncertainty and part (b) accounting for the simulation estimation uncertainty. Therefore, the CI built in Step (5) quantifies the overall uncertainty for $\mu(\mathcal{M}^c)$ estimation.

(1) Identify a design space E for the moment-based parameters $\mathcal{M}$ over which to fit the metamodel. Since the metamodel is used to propagate the input uncertainty measured by the bootstrapped moments $\widetilde{\mathcal{M}}_m$ to the output mean, the design space is chosen to be the smallest ellipsoid covering the most likely bootstrapped moments.
(2) Use a maximin distance Latin hypercube design (LHD) to embed K design points into the design space E. Assign equal replications to K design points to exhaust the simulation budget N and obtain an experiment design $\mathcal{D} = \{(\mathcal{M}^{(i)}, n), i = 1, 2, \ldots, K\}$.
(3) At K design points, generate samples of $\mathbf{X}$ by using the approaches described in Section 4.2. Use these samples to drive simulations and obtain outputs $\mathbf{y}_{\mathcal{D}}$. Compute the sample average $\bar{y}(\mathcal{M}^{(i)})$ and sample variance $s^2(\mathcal{M}^{(i)})$ of the simulation outputs, $i = 1, 2, \ldots, K$. Fit an SK metamodel to obtain $\widehat{m}_p(\cdot)$ and $\widehat{\sigma}_p^2(\cdot)$ using $(\bar{y}(\mathcal{M}^{(i)}), s^2(\mathcal{M}^{(i)}), \mathcal{M}^{(i)})$, $i = 1, 2, \ldots, K$; see Section 4.3.
(4) For $b = 1$ to B:
    (a) Generate bootstrap moments $\widetilde{\mathcal{M}}_m^{(b)}$ by following the procedure in Section 4.1.
    (b) Draw $\widehat{M}_b \sim \mathrm{N}(\widehat{m}_p(\widetilde{\mathcal{M}}_m^{(b)}), \widehat{\sigma}_p^2(\widetilde{\mathcal{M}}_m^{(b)}))$.

Next b:

(5) Report CI: $[\widehat{M}_{(\lceil B\frac{\alpha}{2}\rceil)}, \widehat{M}_{(\lceil B(1-\frac{\alpha}{2})\rceil)}]$, where, $\widehat{M}_{(1)} \le \widehat{M}_{(2)} \le \cdots \le \widehat{M}_{(B)}$ are the sorted values.

To construct the design space E for the SK metamodel, we first generate a test set of bootstrapped moments, denoted by D_T, by following the procedure described in Section 4.1. Then, we find the smallest ellipsoid E that can cover the most likely bootstrapped moments, say, 99%. The ellipsoid's center and shape are the sample mean and covariance matrix of the elements in D_T. The size of D_T is determined by a hypothesis test. See Barton et al. [2014] for more detailed information. By Theorem 4.1, as $m \to \infty$, we have a consistent moment estimator $\widetilde{\mathcal{M}}_m \overset{a.s.}{\to} \mathcal{M}^c$. Thus, the design space E automatically shrinks to a smaller and smaller region around $\mathcal{M}^c$. When m is large enough, by Theorems 4.3 and 4.2, any $\mathcal{M}$ in the design space E eventually has either a feasible NORTA representation or multivariate parametric joint distribution.

The CI $[\widehat{M}_{(\lceil B\frac{\alpha}{2}\rceil)}, \widehat{M}_{(\lceil B(1-\frac{\alpha}{2})\rceil)}]$ provided by our framework characterizes the impact from both input and metamodel uncertainty on a system performance estimate. A variance decomposition in Xie et al. [2015] 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 do both activities to improve the estimation accuracy of $\mu(\mathcal{M}^c)$.

If SK parameters $(\tau^2, \boldsymbol{\zeta}, C)$ are known and we replace $\widehat{M}_b$ in Step (4.b) of the CI procedure with $M_b \sim \mathrm{N}(m_p(\widetilde{\mathcal{M}}_m^{(b)}), \sigma_p^2(\widetilde{\mathcal{M}}_m^{(b)}))$, the CI obtained is $[M_{(\lceil B\frac{\alpha}{2}\rceil)}, M_{(\lceil B(1-\frac{\alpha}{2})\rceil)}]$. Theorem 4.4 shows that $[M_{(\lceil B\frac{\alpha}{2}\rceil)}, M_{(\lceil B(1-\frac{\alpha}{2})\rceil)}]$ is asymptotically consistent.

THEOREM 4.4. *Suppose conditions for Theorems 4.1, 4.3, and 4.2 and the following additional assumptions hold.*

(1) $\epsilon_j(\mathbf{x}) \overset{i.i.d.}{\sim} N(0, \sigma_\epsilon^2(\mathbf{x}))$ *for any* $\mathbf{x}$*, and* $M(\mathbf{x})$ *is a stationary, separable GP with a continuous correlation function satisfying*

$$1 - \gamma(\mathbf{x} - \mathbf{x}') \le \frac{c}{|log(\| \mathbf{x} - \mathbf{x}' \|_2)|^{1+\delta_1}} \text{ for all } \| \mathbf{x} - \mathbf{x}' \|_2 \le \delta_2 \tag{13}$$

for some $c > 0$, $\delta_1 > 0$, *and* $\delta_2 < 1$, *where* $\| \mathbf{x} - \mathbf{x}' \|_2 = \sqrt{\sum_{j=1}^{d'} (x_j - x'_j)^2}$.

(2) The input processes, simulation noise $\epsilon_j(\mathbf{x})$, *and GP* $M(\mathbf{x})$ *are mutually independent and the bootstrap process is independent of all of them.*

Then the interval $[M_{(\lceil B\frac{\alpha}{2}\rceil)}, M_{(\lceil B(1-\frac{\alpha}{2})\rceil)}]$ *is asymptotically consistent, meaning the iterated limit*

$$\lim_{m\to\infty} \lim_{B\to\infty} \Pr\{M_{(\lceil B\alpha/2\rceil)} \le M_p(\mathcal{M}^c) \le M_{(\lceil B(1-\alpha/2)\rceil)}\} = 1 - \alpha. \tag{14}$$

The detailed proof of Theorem 4.4 is provided in the Online Appendix.

Remark 4.5. Theorem 4.4 is based on the assumption that the Gaussian process can correctly characterize the estimation uncertainty of the underlying true response surface given the prior information on $\mu(\cdot)$ and the information obtained from the simulation experiments. The interval $[M_{(\lceil B\frac{\alpha}{2}\rceil)}, M_{(\lceil B(1-\frac{\alpha}{2})\rceil)}]$ constructed by our framework is a CI in the frequentist sense. SK reduces the uncertainty about $\mu(\mathcal{M}^c)$ by simulating at a set of design points. The conditional distribution of $M(\cdot)$ given simulation outputs at design points $\mathbf{Y}_\mathcal{D}$ allows more precise inference about $\mu(\mathcal{M}^c)$. In SK, the distribution $M(\cdot)|\mathbf{Y}_\mathcal{D}$ characterizes the remaining uncertainty about $\mu(\cdot)$. *The interval*

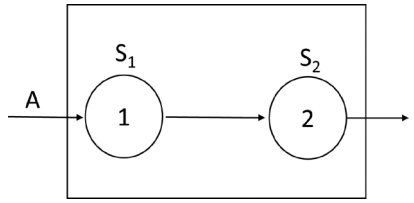

Fig. 1. A queueing system.

$[M_{(\lceil B\frac{\alpha}{2}\rceil)}, M_{(\lceil B(1-\frac{\alpha}{2})\rceil)}]$ *in Equation (14) could be considered as an interval to cover* $M(\mathcal{M}^c)$ *since* $\lim_{m\to\infty}\lim_{B\to\infty}\int \Pr\{M(\mathcal{M}^c) \in [M_{(\lceil B\frac{\alpha}{2}\rceil)}, M_{(\lceil B(1-\frac{\alpha}{2})\rceil)}]|\mathbf{Y}_{\mathcal{D}}\}dP_{\mathbf{Y}_{\mathcal{D}}} = 1-\alpha$; *it is interpreted as a CI for* $\mu(\mathcal{M}^c)$ *because SK treats* $\mu(\cdot)$ *as a realization of* $M(\cdot)$, where $\mathrm{P}_{\mathbf{Y}_{\mathcal{D}}}$ denotes the joint distribution of simulation responses $\mathbf{Y}_{\mathcal{D}}$. Notice that this asymptotic consistency is different from the typical asymptotically valid coverage in the simulation studies that considers the asymptotic performance as the simulation budget goes to infinity.

The metamodel-assisted bootstrapping builds a metamodel based on the simulation results at well-chosen design points and uses the metamodel to propagate the input uncertainty quantified by the bootstrapped samples, $\widetilde{\mathcal{M}}_m^{(b)}$ with $b = 1, 2, \ldots, B$, to the output mean. To run simulations, we need to construct feasible joint input distributions at each design point. When the parametric distribution for F is unknown, some bootstrapped moments may be NORTA infeasible. Therefore, for finite real-world data, the design space E built to cover the most likely bootstrapped samples could include moment combinations $\mathcal{M}$ that are also NORTA infeasible. However, when the dimension of the correlated random vector is relatively low, say, $d \leq 5$, and the pairwise correlation is not so extreme or close to ± 1, the NORTA infeasible problem typically is not an issue for the sample sizes of real-world data encountered in many applications. Therefore, we remove any NORTA infeasible design points in the design space E in Step (2) and assign equal replications to the remaining points. Then, we construct an SK metamodel and use it to estimate mean responses at all bootstrapped samples $\widetilde{\mathcal{M}}_m^{(b)}$ with $b = 1, 2, \ldots, B$. The experiment results in Section 5 indicate that directly throwing away the NORTA infeasible design points does not have an obvious impact on the performance of the metamodel-assisted bootstrapping approach. Notice that since the bootstrapped samples of input moments converge to $\mathcal{M}^c$ and the ellipsoid design space also shrinks around $\mathcal{M}^c$, an interior point of the NORTA feasible region, the NORTA infeasibility could be reduced by obtaining more real-world data for the input models.

5. EMPIRICAL STUDY

In this section, we use the queueing system in Figure 1 to examine the finite-sample performance of our metamodel-assisted bootstrapping approach. Starting with an empty system, we are interested in the expected number of customers in the system over a time interval $[0, T]$ with $T = 100$ time units. The interarrival times follow an exponential distribution, $A \sim \exp(\lambda)$, and the service times at stations 1 and 2 also follow exponential distributions $S_1 \sim \exp(\mu_1)$ and $S_2 \sim \exp(\mu_2)$. There exists component-wise dependence in $\mathbf{X} = (A, S_1, S_2)$.

We assume that F^c is NORTA. For the marginal distributions, the arrival rate $\lambda^c = 1$ and the service rates $\mu_1^c = \mu_2^c = 1.2$. We consider two cases with dependence measured by either Spearman rank or product-moment correlations. The true correlation

Table I. Estimated System Mean Response

$\rho^c_{\mathbf{X}} = Q$	$\rho^c = 0$	$\rho^c = 0.2$	$\rho^c = 0.4$	$\rho^c = 0.6$	$\rho^c = 0.8$
Estimated $\mu(\boldsymbol{\vartheta}^c)$ mean	6.921	6.411	5.913	5.411	4.96
Estimated $\mu(\boldsymbol{\vartheta}^c)$ SE	0.01	0.009	0.007	0.0054	0.0036
$R^c_{\mathbf{X}} = Q$	$\rho^c = 0$	$\rho^c = 0.2$	$\rho^c = 0.4$	$\rho^c = 0.6$	$\rho^c = 0.8$
Estimated $\mu(\boldsymbol{\vartheta}^c)$ mean	6.93	6.475	5.977	5.472	4.993
Estimated $\mu(\boldsymbol{\vartheta}^c)$ SE	0.01	0.009	0.007	0.006	0.004

matrices $R^c_{\mathbf{X}}$ or $\rho^c_{\mathbf{X}}$ are

$$Q \equiv \begin{pmatrix} 1 & \rho^c & \rho^c \\ & 1 & \rho^c \\ & & 1 \end{pmatrix}.$$

For illustration, we set off-diagonal elements in the correlation matrices as a constant value. Therefore, the number of parameters characterizing the input model F is $d' = d^\dagger + d^* = 3 + (3 \times 2)/2 = 6$. Since the true mean response $\mu(\boldsymbol{\vartheta}^c)$ is unknown, we run 10^5 replications to estimate it with results shown in Table I, which records the mean and standard error (SE) of the estimated system response for different values of ρ^c. Notice that the dependence level quantified by ρ^c significantly impacts the system mean response. In this section, we present the empirical results for $\rho^c = 0.4$ or 0.8, which are representative of the performance of our metamodel-assisted bootstrapping approach.

To evaluate metamodel-assisted bootstrapping, we pretend that the input model parameters $(\boldsymbol{\theta}^c, R^c_{\mathbf{X}})$ or $(\boldsymbol{\theta}^c, \rho^c_{\mathbf{X}})$ are unknown and that they are estimated by m i.i.d. observations from F^c; this represents obtaining "real-world data." The goal is to build a CI quantifying the impact of both input and simulation estimation error on the system mean response estimate.

We compare metamodel-assisted bootstrapping to the conditional CI and direct bootstrapping. For the conditional CI, we fit the input distribution to the real-world data by moment matching and allocate the entire computational budget of N replications to simulating the resulting system. In direct bootstrapping, we run N/B replications of the simulation at each bootstrap moment $\widetilde{\mathcal{M}}^{(b)}_m$, record the average simulation output $\bar{Y}_b = \bar{Y}(\widetilde{\mathcal{M}}^{(b)}_m)$, and report the percentile CI $[\bar{Y}_{(\lceil B\frac{\alpha}{2}\rceil)}, \bar{Y}_{(\lceil B(1-\frac{\alpha}{2})\rceil)}]$. In metamodel-assisted bootstrapping, we evenly assign N replications to K design points, run simulations, build an SK metamodel, and record the percentile CI $[\widehat{M}_{(\lceil B\frac{\alpha}{2}\rceil)}, \widehat{M}_{(\lceil B(1-\frac{\alpha}{2})\rceil)}]$ by following the procedure in Section 4.4.

When we construct the stochastic kriging metamodel, we first use an LHD to find potential design points to evenly cover the ellipsoid design space E. There may exist NORTA infeasible design points. For ease of implementation, we throw away those NORTA infeasible points and allocate all the computational budget to the remaining design points, which is called "Design D1." This experiment design would not cause metamodel bias under the assumption that the true unknown response surface $\mu(\cdot)$ is a realization of a GP. To see if directly removing the NORTA infeasible design points could impact the performance of our metamodel-assisted bootstrap approach, we also make some modification and obtain "Design D2." Since Ghosh and Henderson [2002b] indicate that we could always find a *close* NORTA feasible approximation to a NORTA infeasible point, we replace NORTA infeasible points with close new design points that are NORTA feasible. Specifically, we find a close positive semidefinite approximation for $\rho_{\mathbf{Z}}$, denoted by $\bar{\rho}_{\mathbf{Z}}$ [Higham 2002], and let $\rho^a_{\mathbf{Z}} \equiv \bar{\rho}_{\mathbf{Z}} + \delta' \mathrm{I}_{d\times d}$, where $\mathrm{I}_{d\times d}$ denotes a $d \times d$ identity matrix and δ' is a small positive value. We use $\delta' = 10^{-7}$ in the empirical

Table II. Results for Nominal 95% CIs When $m = 100, 500, 1,000$ When the Dependence Is Characterized by Spearman Rank Correlations $R^c_{\mathbf{X}} = Q$

$m = 100$		$\rho^c = 0.4$		$\rho^c = 0.8$	
		$N = 10^3$	$N = 10^4$	$N = 10^3$	$N = 10^4$
conditional CI	coverage	5.2%	2.7%	7.6%	3.4%
	CI width (mean)	0.322	0.099	0.158	0.052
	CI width (SD)	0.117	0.034	0.043	0.014
direct bootstrap	coverage	99.4%	96.3%	100%	96.8%
	CI width (mean)	14.301	9.515	6.778	4.402
	CI width (SD)	4.588	3.214	1.908	1.469
metamodel-assisted bootstrap	coverage	93.8%	94.2%	95.8%	95.1%
	CI width (mean)	9.248	8.954	3.978	4.095
	CI width (SD)	3.43	3.084	1.307	1.393
$m = 500$		$\rho^c = 0.4$		$\rho^c = 0.8$	
		$N = 10^3$	$N = 10^4$	$N = 10^3$	$N = 10^4$
conditional CI	coverage	12.3%	4%	15.5%	4.9%
	CI width (mean)	0.298	0.094	0.151	0.048
	CI width (SD)	0.053	0.017	0.019	0.006
direct bootstrap	coverage	100%	99.1%	100%	99.2%
	CI width (mean)	10.415	4.672	5.172	2.196
	CI width (SD)	1.827	0.863	0.714	0.312
metamodel-assisted bootstrap	coverage	94.7%	94.4%	96.1%	94.6%
	CI width (mean)	3.587	3.518	1.562	1.52
	CI width (SD)	0.795	0.693	0.3	0.246
$m = 1000$		$\rho^c = 0.4$		$\rho^c = 0.8$	
		$N = 10^3$	$N = 10^4$	$N = 10^3$	$N = 10^4$
conditional CI	coverage	17.4%	6.9%	24.4%	7.8%
	CI width (mean)	0.295	0.093	0.149	0.047
	CI width (SD)	0.039	0.011	0.014	0.004
direct bootstrap	coverage	100%	99.9%	100%	99.9%
	CI width (mean)	9.881	3.911	4.968	1.874
	CI width (SD)	1.265	0.493	0.504	0.184
metamodel-assisted bootstrap	coverage	95.3%	94.5%	95.2%	94.7%
	CI width (mean)	2.626	2.439	1.171	1.044
	CI width (SD)	0.532	0.345	0.236	0.123

study. Then, we set $\rho^a_{\mathbf{Z}}$ as the correlation matrix for NORTA and generate samples of $\mathbf{X}$ for simulation runs.

In direct bootstrapping, for the small percentage of NORTA infeasible bootstrap resampled moments $\widetilde{\mathcal{M}}^{(b)}_m$, we first find a close NORTA feasible approximation by following the approach used in Design D2. Then, we use this approximated input model to drive the simulations and estimate the system performance.

For the input distribution with dependence characterized by either Spearman rank or product-moment correlations, Tables II and III show the statistical performance of conditional and direct bootstrapping CIs and metamodel-assisted bootstrapping with $m = 100, 500, 1{,}000$ real-world observations and computational budget of $N = 10^3, 10^4$ replications. Since the studies by Jones et al. [1998] and Loeppky et al. [2009] recommend that the number of design points should be 10 times the dimension of the problem for kriging, we set the number of design points $K = 60$. We ran 1,000 macro-replications of the entire experiment. In each macro-replication, we first generate m multivariate observations by using NORTA with parameters $(\boldsymbol{\theta}^c, R^c_{\mathbf{X}})$ or $(\boldsymbol{\theta}^c, \rho^c_{\mathbf{X}})$. Then,

Table III. Results for Nominal 95% CIs When $m = 100, 500, 1{,}000$ When the Dependence Is Characterized by Product-Moment Correlations $\rho^c_{\mathbf{X}} = Q$

$m = 100$		$\rho^c = 0.4$		$\rho^c = 0.8$	
		$N = 10^3$	$N = 10^4$	$N = 10^3$	$N = 10^4$
conditional CI	coverage	7.7%	2.1%	7.6%	2%
	CI width (mean)	0.315	0.097	0.155	0.049
	CI width (SD)	0.117	0.034	0.045	0.014
direct bootstrap	coverage	99.1%	97.7%	99.8%	96.2%
	CI width (mean)	13.87	9.279	6.448	4.072
	CI width (SD)	4.53	3.074	1.922	1.398
metamodel-assisted bootstrap (Design D1)	coverage	93.3%	95.6%	95.4%	95.4%
	CI width (mean)	9.075	8.754	3.844	3.819
	CI width (SD)	3.468	3.074	1.372	1.338
metamodel-assisted bootstrap (Design D2)	coverage	93.2%	95.6%	95.4%	95.3%
	CI width (mean)	9.08	8.754	3.841	3.812
	CI width (SD)	3.466	3.073	1.388	1.335
$m = 500$		$\rho^c = 0.4$		$\rho^c = 0.8$	
		$N = 10^3$	$N = 10^4$	$N = 10^3$	$N = 10^4$
conditional CI	coverage	14.4%	4.8%	14.9%	6.6%
	CI width (mean)	0.287	0.09	0.145	0.046
	CI width (SD)	0.051	0.016	0.02	0.006
direct bootstrap	coverage	100%	98.8%	100%	99.2%
	CI width (mean)	9.742	4.42	4.819	2.051
	CI width (SD)	1.703	0.814	0.682	0.289
metamodel-assisted bootstrap (Design D1)	coverage	94.2%	94.7%	94.6%	95%
	CI width (mean)	3.419	3.386	1.45	1.435
	CI width (SD)	0.751	0.671	0.266	0.23
metamodel-assisted bootstrap (Design D2)	coverage	94.5%	95%	95.3%	94.9%
	CI width (mean)	3.421	3.387	1.45	1.435
	CI width (SD)	0.749	0.669	0.27	0.231
$m = 1000$		$\rho^c = 0.4$		$\rho^c = 0.8$	
		$N = 10^3$	$N = 10^4$	$N = 10^3$	$N = 10^4$
conditional CI	coverage	20.4%	6.6%	20.7%	6.2%
	CI width (mean)	0.282	0.089	0.145	0.046
	CI width (SD)	0.035	0.011	0.014	0.004
direct bootstrap	coverage	100%	99.9%	100%	100%
	CI width (mean)	9.183	3.664	4.644	1.766
	CI width (SD)	1.135	0.485	0.467	0.182
metamodel-assisted bootstrap (Design D1)	coverage	95.5%	94.8%	95.1%	93.9%
	CI width (mean)	2.450	2.328	1.088	1.001
	CI width (SD)	0.449	0.332	0.194	0.125
metamodel-assisted bootstrap (Design D2)	coverage	95.5%	94.7%	95.2%	93.9%
	CI width (mean)	2.454	2.328	1.086	1.002
	CI width (SD)	0.449	0.332	0.192	0.124

for the conditional CI, we run N replications at the estimated parameters $(\widehat{\boldsymbol{\theta}}_m, \widehat{R}_{\mathbf{X},m})$ or $(\widehat{\boldsymbol{\theta}}_m, \widehat{\rho}_{\mathbf{X},m})$ and build CIs with nominal 95% coverage of the response mean. For direct bootstrapping and metamodel-assisted bootstrapping, we use bootstrapping to generate $B = 1{,}000$ sample moments to quantify the input uncertainty. Since $\mu(\cdot)$ is unknown, we use the fixed computational budget N to propagate the input uncertainty either via direct simulation or via the SK metamodel to build percentile CIs with nominal 95% coverage.

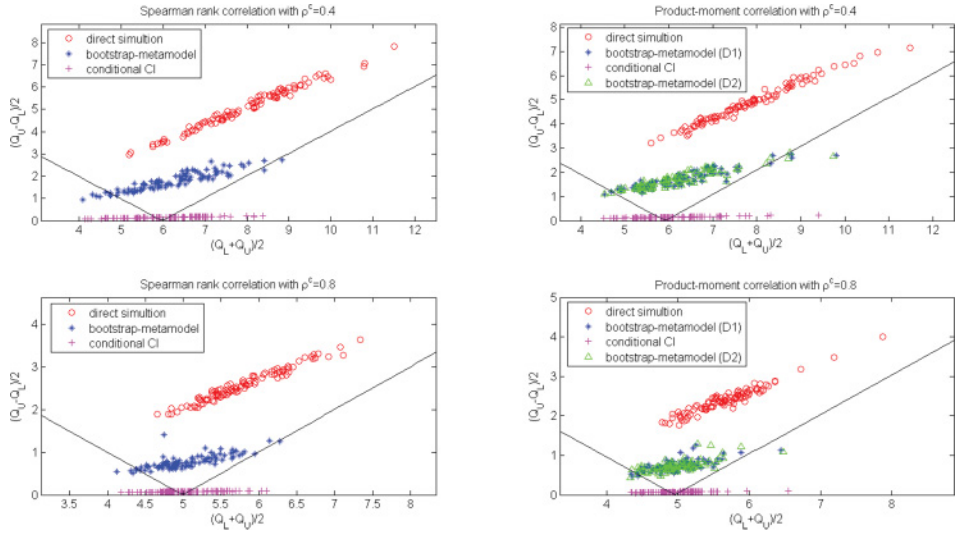

Fig. 2. Scatter plots of conditional CIs and CIs obtained by direct bootstrapping and metamodel-assisted bootstrapping with $m = 500$, $N = 10^3$, and $\rho^c = 0.4, 0.8$.

When we use the product-moment correlation to measure the dependence, the probability to get NORTA infeasible bootstrapped moments $\widetilde{\mathcal{M}}_m^{(b)}$ is negligible when $\rho^c = 0.4$. For $\rho^c = 0.8$, the probability has mean 3.55% and standard deviation 2.65% when $m = 100$, mean 3.14% and standard deviation 4.08% when $m = 500$, and mean 1.54% and standard deviation 3.29% when $m = 1{,}000$. When we use the Spearman rank correlation to measure the dependence, the probability to get NORTA infeasible bootstrapped moments is negligible when $\rho^c = 0.4, 0.8$.

From Tables II and III, the results with dependence measured by either product-moment or Spearman rank correlations are similar. We observe that under the same computational budget N, the conditional CIs that only account for the simulation uncertainty tend to have undercoverage. The CIs obtained by direct bootstrapping are much wider and they typically have obvious overcoverage. The CIs obtained by metamodel-assisted bootstrapping with Designs D1 and D2 have similar performance and they have coverage much closer to the nominal level of 95%. As N increases and the simulation estimation error decreases, the undercoverage problem for the conditional CI becomes worse. Since direct bootstrap and metamodel-assisted bootstrap use the same set of bootstrapped samples to quantify input uncertainty, the overcoverage for the direct bootstrap represents the additional simulation uncertainty introduced while propagating the input uncertainty to the output mean. Tables II and III show that the metamodel can effectively use the computational budget and reduce the impact from the simulation estimation error. Further, as the computational budget increases, the difference between the CIs obtained by the two methods diminishes.

Figure 2 shows scatter plots of conditional CIs and CIs obtained by direct bootstrapping and metamodel-assisted bootstrapping with $m = 500$, $N = 10^3$, and $\rho^c = 0.4, 0.8$ when we use either Spearman-rank or product-moment correlations. They include results from 100 macro-replications. The horizontal axis represents $(Q_L + Q_U)/2$, the center of the CI, where Q_L and Q_U are the lower and upper bounds of the CIs. The vertical axis is $(Q_U - Q_L)/2$, the half-width of the CIs. Region 1 contains points that correspond to CIs having underestimation and Region 3 contains points corresponding to overestimation, while Region 2 contains CIs that cover $\mu(F^c)$; see Kang and Schmeiser

[1990]. The conclusions obtained for Spearman rank and product-moment correlations are similar. Since the standard deviation of the system response estimate increases with the mean, we observe that all CIs tend to be wider when the center is larger. Conditional CIs have width too short and their centers have large variance. Therefore, they have serious undercoverage. The variance for centers of CIs comes from the impact of input uncertainty. Since metamodel-assisted bootstrapping accounts for both input and simulation uncertainty, its CI width is large enough to avoid undercoverage. The proportion of CIs in Region 2 is close to 95%, and CIs outside tend to have underestimation based on results from 1,000 macro-replications. The CIs obtained by Designs D1 and D2 have similar performance. The width of CIs obtained by the direct bootstrap is too large; all CIs are located in Region 2 and they have serious overcoverage.

6. CONCLUSIONS

In this paper, we extended the metamodel-assisted bootstrapping framework of Xie et al. [2015] to stochastic simulation with dependent input models. The input models are characterized by their marginal distribution parameters and dependence measured either by Spearman rank or product-moment correlations, which are estimated from real-world data. Metamodel-assisted bootstrapping uses the bootstrap to quantify the estimation error of these joint distributions and propagates it to the output mean by using an equation-based SK metamodel. We proposed a procedure to deliver a CI quantifying the overall uncertainty of the system performance estimate. The asymptotic consistency of this interval is proved under the assumption that the true mean response surface is a realization of a GP. Our metamodel-assisted bootstrap framework is applicable to cases when the parametric family of multivariate input distribution is known or unknown. When the parametric joint input distributions are unknown, we construct the joint distributions by using the flexible NORTA representation.

An empirical study using a queueing example demonstrates that for the input distribution with dependence measured by either Spearman rank or product-moment correlations, our metamodel-assisted bootstrap approach has good finite-sample performance under various quantities of real-world data and simulation budget. When the simulation budget is tight, compared with the direct bootstrap, the metamodel-assisted bootstrap can make more effective use of the simulation budget.

ELECTRONIC APPENDIX

The electronic appendix for this article can be accessed in the ACM Digital Library.

ACKNOWLEDGMENTS

This research was partially supported by National Science Foundation Grant CMMI-1068473 and GOALI sponsor Simio. The authors thank Cheng Li in the proof of Lemma A.3. Portions of this were previously published in the *Proceedings of the 2014 Winter Simulation Conference* as Xie et al. [2014b].

REFERENCES

Bruce E. Ankenman, Barry L. Nelson, and Jeremy Staum. 2010. Stochastic kriging for simulation metamodeling. *Operations Research* 58 (2010), 371–382.

Eusebio Arenal-Gutiérrez, Carlos Matrán, and Juan A. Cuesta-Albertos. 1996. Unconditional Glivenko-Gantelli-type theorems and weak laws of large numbers for bootstrap. *Statistics & Probability Letters* 26 (1996), 365–375.

Francois Bachoc. 2013. Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model misspecification. *Computational Statistics & Data Analysis* 66 (2013), 55–69.

Russell R. Barton. 2007. Presenting a more complete characterization of uncertainty: Can it be done? In *Proceedings of the 2007 INFORMS Simulation Society Research Workshop*. INFORMS Simulation Society, Fontainebleau.

Russell R. Barton. 2012. Tutorial: Input uncertainty in output analysis. In *Proceedings of the 2012 Winter Simulation Conference*, C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher (Eds.). IEEE Computer Society, 67–78.

Russell R. Barton, Barry L. Nelson, and Wei Xie. 2014. Quantifying input uncertainty via simulation confidence intervals. *Informs Journal on Computing* 26 (2014), 74–87.

Russell R. Barton and Lee W. Schruben. 1993. Uniform and bootstrap resampling of input distributions. In *Proceedings of the 1993 Winter Simulation Conference*, G. W. Evans, M. Mollaghasemi, E. C. Russell, and W. E. Biles (Eds.). IEEE Computer Society, 503–508.

Russell R. Barton and Lee W. Schruben. 2001. Resampling methods for input modeling. In *Proceedings of the 2001 Winter Simulation Conference*, B. A. Peters, J. S. Smith, D. J. Medeiros, and M. W. Rohrer (Eds.). IEEE Computer Society, 372–378.

Bahar Biller and Canan G. Corlu. 2011. Accounting for parameter uncertainty in large-scale stochastic simulations with correlated inputs. *Operations Research* 59 (2011), 661–673.

Bahar Biller and Soumyadip Ghosh. 2006. Multivariate input processes. In *Handbooks in Operations Research and Management Science: Simulation*, S. Henderson and B. L. Nelson (Eds.). Elsevier, Chapter 5.

Patrick Billingsley. 1995. *Probability and Measure*. Wiley-Interscience, New York.

Marne C. Cario and Barry L. Nelson. 1997. *Modeling and Generating Random Vectors with Arbitrary Marginal Distributions and Correlation Matrix*. Technical report. Department of Industrial Engineering and Management Sciences, Northwestern University.

Xi Chen, Bruce E. Ankenman, and Barry L. Nelson. 2012. The effect of common random numbers on stochastic kriging metamodels. *ACM Transactions on Modeling and Computer Simulation* 22 (2012), 7:1–7:20.

Russell C. H. Cheng and Wayne Holland. 1997. Sensitivity of computer simulation experiments to errors in input data. *Journal of Statistical Computation and Simulation* 57 (1997), 219–241.

Robert T. Clemen and Terence Reilly. 1999. Correlations and copulas for decision and risk analysis. *Management Science* 45 (1999), 208–224.

Sourav Das, Tata S. Rao, and Georgi N. Boshnakov. 2012. *On the Estimation of Parameters of Variograms of Spatial Stationary Isotropic Random Processes*. Research Report No. 2. The University of Manchester.

Soumyadip Ghosh and Shane G. Henderson. 2002a. Chessboard distributions and random vectors with specified marginals and covariance matrix. *Operations Research* 50 (2002), 820–834.

Soumyadip Ghosh and Shane G. Henderson. 2002b. Properties of the NORTA method in higher dimensions. In *Proceedings of the 2002 Winter Simulation Conference*, E. Yűcesan, C. H. Chen, J. L. Snowdon, and J. M. Charnes (Eds.). IEEE Computer Society, 263–269.

Peter Hall. 1988. Rate of convergence in bootstrap approximations. *The Annals of Probability* 16 (1988), 1665–1684.

Shane G. Henderson, Belinda A. Chiera, and Roger M. Cooke. 2000. Generating dependent quasi-random numbers. In *Proceedings of the 2000 Winter Simulation Conference*, J. A. Joines, R. R. Barton, K. Kang, and P. A. Fishwick (Eds.). IEEE Computer Society, 527–536.

Nicholas J. Higham. 2002. Computing the nearest correlation matrix – A problem from finance. *IMA Journal on Numerical Analysis* 22 (2002), 329–343.

Mark E. Johnson. 1987. *Multivariate Statistical Simulation*. Wiley, New York.

Donald R. Jones, Matthias Schonlau, and William J. Welch. 1998. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization* 13 (1998), 455–492.

Keebom Kang and Bruce Schmeiser. 1990. Graphical methods for evaluation and comparing confidence-interval procedures. *Operations Research* 38, 3 (1990), 546–553.

Shing T. Li and Joseph L. Hammond. 1975. Generation of pseudorandom numbers with specified univariate distributions and correlation coefficients. *IEEE Transactions on Systems, Man, and Cybernetics* 5 (September 1975), 557–561.

Jason L. Loeppky, Jerome Sacks, and William J. Welch. 2009. Choosing the sample size of a computer experiment: A practical guide. *Technometrics* 51 (2009), 366–376.

Tanmoy Mukhopadhyay, Sushanta Chakroborty, Sondipon Adhikari, and Rajib Chowdhury. 2016. A critical assessment of kriging model variants for high-fidelity uncertainty quantification in dynamics of composite shells. *Archives on Computational Methods in Engineering* (2016). Online version.

Bruce W. Schmeiser and Ram Lal. 1982. Bivariate gamma random vectors. *Operations Research* 30, 2 (1982), 355–374.

Jun Shao and Dongsheng Tu. 1995. *The Jackknife and Bootstrap*. Springer-Verlag.

Eunhye Song, Barry L. Nelson, and C. Dennis Pegden. 2014. Advanced tutorial: Input uncertainty quantification. In *Proceedings of the 2014 Winter Simulation Conference*, A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller (Eds.). IEEE Computer Society.

A. W. Van Der Vaart. 1998. *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK.

William R. Wade. 2010. *An Introduction to Analysis* (4th ed.). Prentice Hall.

Wei B. Wu and Jan Mielniczuk. 2010. A new look at measuring dependence. In *Dependence in Probability and Statistics*, P. Doukhan, G. Lang, D. Surgailis, and G. Teyssière (Eds.). Springer.

Wei Xie, Barry L. Nelson, and Russell R. Barton. 2014a. A Bayesian framework for quantifying uncertainty in stochastic simulation. *Operational Research* 62, 6 (2014), 1439–1452.

Wei Xie, Barry L. Nelson, and Russell R. Barton. 2014b. Statistical uncertainty analysis for stochastic simulation with dependent input models. In *Proceedings of the 2014 Winter Simulation Conference*, A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller (Eds.). IEEE Computer Society.

Wei Xie, Barry L. Nelson, and Russell R. Barton. 2015. Statistical uncertainty analysis for stochastic simulation. (2015). Working Paper, Department of Industrial and Systems Engineering, Rensselaer Polytechnic Institute, Troy, NY.

Wei Xie, Barry L. Nelson, and Jeremy Staum. 2010. The influence of correlation functions on stochastic kriging metamodels. In *Proceedings of the 2010 Winter Simulation Conference*, B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yucesan (Eds.). IEEE Computer Society, 1067–1078.

Jun Yin, Szu H. Ng, and Kien M. Ng. 2009. A study on the effects of parameter estimation on kriging model's prediction error in stochastic simulation. In *Proceedings of the 2009 Winter Simulation Conference*, B. Johansson A. Dunkin M. D. Rossetti, R. R. Hill and R. G. Ingalls (Eds.). IEEE Computer Society, 674–685.

Received October 2014; revised July 2016; accepted August 2016