Submitted to *Operations Research*
manuscript (Please, provide the manuscript number!)

A Bayesian Nonparametric Hierarchical Framework for Uncertainty Quantification in Simulation

Wei Xie
Department of Industrial and Systems Engineering, Rensselaer Polytechnic Institute, xiew3@rpi.edu
Cheng Li
Department of Statistics and Applied Probability, National University of Singapore, stalic@nus.edu.sg
Pu Zhang
Department of Industrial and Systems Engineering, Rensselaer Polytechnic Institute, zhangp5@rpi.edu

When we use the simulation to assess the performance of stochastic systems, the input models used to drive simulation experiments are often estimated from finite real-world data. There exist both input and simulation estimation uncertainty in the system performance estimates. Without strong prior information on the input models and the system mean response surface, in this paper, we propose a Bayesian nonparametric hierarchical framework to quantify the impact from both sources of uncertainty. Specifically, nonparametric input models are introduced to faithfully capture the important features of the real-world data, such as heterogeneity, multi-modality, skewness and heavy tails. Bayesian posteriors of flexible input models characterize the input uncertainty, which automatically accounts for both model selection and parameter value uncertainty. Then, the input uncertainty is propagated to outputs by using the direct simulation with simulation estimation uncertainty quantified by the posterior distribution of the mean system response. Thus, under very general conditions, our framework delivers a credible interval accounting for both input and simulation uncertainty. A variance decomposition is further developed to quantify the relative contributions from both sources of uncertainty. Our approach is supported by rigorous theoretical and empirical study.

Key words: Bayesian nonparametric model; stochastic simulation; input modeling; input uncertainty; uncertainty quantification

History:

1. Introduction

Simulation is widely used in many applications to assess the performance of stochastic systems, e.g., manufacturing, supply chain and health care systems. The input models, defined as the driving stochastic processes in simulation experiments, are often estimated from finite real-world data. Therefore, there exist both input estimation error, called *the input uncertainty*, and simulation estimation error, called *the simulation uncertainty*. Ignoring either source of uncertainty could lead to unfounded confidence in the simulation assessment of system performance.

Various approaches have been proposed in the literature to quantify the input and simulation uncertainty; see Barton (2012), Song et al. (2014) and Lam (2016) for a comprehensive review. Based on methodologies developed for characterizing the input uncertainty, they can be divided into frequentist and Bayesian approaches. The frequentist approaches typically study the sampling distributions of point estimators of underlying input models. Since it could be hard to get the exact sampling distributions in many situations, the asymptotic approximation, including the normal approximation and the bootstrap, is often used to quantify the input uncertainty, which is valid when the amount of real-world data is large. *However, even in the current big data world, we often face the situations where the amount of real-world data is limited, especially for the high-tech products with short life cycles.* For example, biopharma manufacturing requires 9 to 12 months from raw materials sourcing to the finish drug products, and it requires another 2 to 3 months for quality testing. However, the drug substances typically expire after 18 to 36 months; see Otto et al. (2014). Compared to frequentist methods, Bayesian approaches derive the posterior distributions quantifying the input uncertainty and they do not need a large-sample asymptotic approximation for their validation. It is also straightforward for Bayesian approaches to incorporate the prior information about the underlying input models. See Xie et al. (2014) for the discussion of frequentist v.s. Bayesian approaches for input uncertainty.

In this paper, we focus on developing a rigorous Bayesian framework to quantify the estimation uncertainty of system mean performance when we do not have strong prior information on the

input models and system mean response surface. We consider univariate input models, which model independent and identically distributed (i.i.d.) data by mutually independent input distributions.

Many existing methods assume specific parametric families for input models with unknown parameter values estimated from finite real-world data. Thus, the input uncertainty can be quantified by the posteriors of input parameters, see for example Cheng and Currie (2003), Ng and Chick (2006) and Xie et al. (2014), etc. Parametric approaches tend to work well when one has strong prior beliefs on the families of input models. However, they are considered to be too restrictive in general. If the selected parametric families do not have sufficient flexibility and cannot represent the underlying input models well, there always exists the distribution family selection error which does not vanish as the amount of real-world data becomes large. This inconsistent estimation could lead to incorrect inference even for the moderate size of real-world data (Hjort et al. 2011).

One possible remedy for the inconsistency of parametric approaches is to introduce the *family uncertainty*, which accounts for the input model selection error among a prespecified pool of different candidate parametric families. For example, Chick (2001) proposed the Bayesian Model Averaging (BMA) to quantify input uncertainty from both families and parameter values, where the family uncertainty is characterized by the posterior probabilities of different candidate parametric models. However, *BMA is based on the assumption that all data come from one of candidate distributions (Bishop 2006).* In other words, BMA relies on the assumption that all data are generated from a *single* true underlying parametric family, and this family must be included as a candidate *priori*. Furthermore, if the selected parametric families are not mutually exclusive, such as exponential and Gamma distributions, it can potentially lead to model identification problems.

Instead of using parametric families, we explore *nonparametric* input modeling in the Bayesian framework that we develop to quantify the overall system performance estimation uncertainty. Our approach is based on the following consideration: *In many situations, the real-world data represent the variability caused by various latent sources of uncertainty.* It is almost impossible to find a single parametric family that can capture the important features in the real-world data. For

example, in a raw material (RM) inventory system in the bio-pharmaceutical manufacturing, a single RM is used to satisfy the demands from various production lines, called *move orders*. Different latent sources of uncertainty in production lines could lead to multi-modality. Heterogeneity is commonly observed because the variability of different sources of uncertainty could be different. In addition, contamination and cross-contamination in the production processes, which could cause the shutdown of production lines and the throwaway of batches of products, can lead to the right skewness and the tail in the move orders. As a result, the underlying physical input distributions characterizing the variability from various sources of uncertainty could have the important features, including heterogeneity, multi-modality, skewness and tails. These important properties are also observed in the real-world data collected from other industries; see for example Wagner et al. (2009), Ma (2011) and Akcay et al. (2011).

Flexible Bayesian nonparametric input models are presented to efficiently capture important features in the real-world data. For discrete random variables with finite support points, the multinomial distribution could be used as a straightforward nonparametric estimator of the true underlying distribution. *Therefore, in this paper, we focus on the input modeling of continuous random variables.* Specifically, our input models are based on the Dirichlet Processes Mixtures (DPM), a popular Bayesian nonparametric modeling method in both statistics and machine learning communities. For details about the DPM, we refer the readers to Ghosh and Ramamoorthi (2003). Motivated by the kernel density estimation (KDE), the Bayesian nonparametric DPM approach with Gaussian kernel was introduced in the statistics community (West 1990, Escobar and West 1995, etc.), which expresses the generative process of real-world data as a nonparametric mixture distribution of normals. It is extended to other kernel functions; see for example Hanson (2006), Kottas (2006), Wu and Ghosal (2008).

From the modeling perspective, DPM has clear advantages over all parametric families because the variability across different mixing components naturally represents various latent sources of uncertainty, which makes it straightforward to capture the important properties in the real-world

data. Different from parametric approaches, the number of active mixing components and parameters can automatically adjust to the complexity of real data. Thus, our empirical study demonstrates that DPM has better and more robust finite sample performance. From the theoretical perspective, DPM is able to consistently estimate a wide class of distributions under very general conditions (Ghosal et al. 1999, Wu and Ghosal 2008, etc.). This generality of consistent estimation is clearly lacking in most parametric methods. Compared to BMA, our approach does not rely on the assumption that the true underlying distribution comes from a particular parametric family, and thus completely avoids the difficulty of select the "appropriate" candidate parametric distributions. From the computational perspective, one can develop efficient posterior samplers for DPM of popular exponential families (see our Section 3.2, Escobar and West 1995, Neal 2000, etc.).

Among frequentist approaches, empirical distribution is the most commonly used nonparametric approach in the simulation literature, and the bootstrap is typically used to quantify the input estimation uncertainty; see for example Barton and Schruben (1993), Barton (2007). Empirical distribution is simple and easy to implement. However, DPM has some important advantages compared to empirical distribution. First, even though the underlying true distribution is continuous, empirical distribution is always discrete. When we have a limited amount of real-world data, samples from the empirical distribution could overlook some important properties in the underlying input models, such as the tails. Second, the validity of using the bootstrap to quantify the input uncertainty relies on large sample asymptotics and therefore requires large samples of real-world data. However, as we mentioned above, the decision makers often face the situations where the amount of real-world data is limited. As a Bayesian approach, DPM can overcome these limitations. Our empirical study demonstrates that DPM has better finite sample performance compared to frequentist competitors, especially when the sample size of real-world data is small. Third, unlike empirical distribution, DPM tries to model the underlying generative processes of inputs, which could be used to identify the latent sources of uncertainty, and further study their impact on the system performance.

Therefore, in this paper, we develop a Bayesian nonparametric hierarchical framework to quantify the system performance estimation uncertainty. We first introduce nonparametric input models based on DPM with various kernels, which could capture the important properties in the real-world data. *The samples drawn from posteriors of flexible input models can automatically quantify both model selection and parameters value uncertainty.* Then, the input uncertainty is propagated to the output through *the direct simulation* that runs simulations at each sample of input models, while the simulation uncertainty is quantified by the posterior distributions of the mean system responses. Our Bayesian framework leads to a sampling procedure that delivers a posterior distribution and further a percentile credible interval (CrI) quantifying the overall uncertainty of system performance estimates.

In sum, the main contributions of our paper are as follows:

1. Considering that the real-world data represent the variability caused by various latent sources of uncertainty in many situations, DPM is used to model the underlying generative processes of inputs. It provides sufficient flexibility to capture the important features in the real-world data, and can overcome the limitations of existing approaches on input uncertainty, including parametric approaches, BMA and empirical distribution. Further, DPM with Gamma, Gaussian and Beta kernels can model input data with support on the non-negative half real line, the whole real line, and an interval with finite length. The empirical study demonstrates that our input models have better and more robust performance than existing approaches.

2. Without prior information about the underlying true input distributions and the system response surface, we propose a Bayesian framework which accounts for both input and simulation uncertainty in the Bayesian paradigm, and also delivers a CrI quantifying the overall uncertainty of system mean performance estimates. Furthermore, a variance decomposition is developed to quantify the relative contributions from the input and simulation uncertainty.

3. We provide a rigorous theoretical support for our Bayesian nonparametric framework. The theory includes the posterior consistency of our CrI accounting for both input and simulation

uncertainty. Given a fixed sample size of the real-world data, as the simulation budget increases, this CrI converge to the CrI accounting for input uncertainty with the true mean response surface known. Further, as the amount of real-world data and the simulation budget go to infinity, the CrI converges to the true system performance.

The next section describes the related studies on input modeling and uncertainty quantification. In Section 3, a Bayesian framework is introduced to quantify the overall uncertainty of the system performance estimates. We then report results of finite sample behaviors on both input and system performance estimation in Section 4, and we conclude this paper in Section 5. All proofs, derivations and other supplementary studies are included in the online Appendix.

2. Background

Since Barton (2012), Song et al. (2014) and Lam (2016) provided the comprehensive review on input uncertainty and uncertainty quantification, in this section, we briefly discuss existing Bayesian approaches related to our approach. When the parametric families of input models are known, samples drawn from posteriors of input parameters can quantify the input uncertainty. Two approaches are typically used to propagate the input uncertainty to the output: direct simulation and metamodeling. Zouaoui and Wilson (2003) run the simulations at each sample of input models. Then, a random effect model and a hierarchical normal model were used to do inference on the system mean response. Both models are built on the homogeneity assumption which requires the constant variance of the simulation error at different posterior samples of input models.

Since each simulation run could be computationally expensive, an equation-based metamodel as a function of input parameters could efficiently propagate the input uncertainty to output. Ng and Chick (2006) developed a first-order metamodel based on a Taylor series approximation. This local approximation is suitable to the situations when there is a large amount of real-world data and posterior distributions locate in a small neighborhood of the true input parameters. To account for more general situations when the amount of real-world data could be small, Xie et al. (2014) built a Gaussian process global metamodel to propagate the input uncertainty to the output mean.

Chick (2001) proposed BMA approach to account for both input model and parameter value uncertainty. Given a set of candidate parametric distributions, Bayesian posteriors of the input families and parameters are derived to quantify the input uncertainty. Then, the simulations are driven by samples from the posteriors of input distributions, and the sample mean of simulation outputs is used to estimate the system posterior mean response. To separate the relative contributions from input model, parameters and simulation uncertainty on the system posterior mean response, Zouaoui and Wilson (2004) developed a BMA-based simulation replication algorithm. The key difference from Chick (2001) is that Zouaoui and Wilson (2004) assigned multiple simulation replications to each sample of input distributions. Confidence intervals (CIs) are constructed to quantify the overall variability for the posterior mean response. BMA is further extended to input models with dependence by using Normal-to-Anything (NORTA) (Biller and Corlu 2011). However, since BMA is based on the assumption that the real-world data come from a single candidate parametric distribution, it could be challenging to choose appropriate candidate distributions when we do not have strong prior information on the underlying input models. In addition, BMA does not quantify the simulation uncertainty in the Bayesian manner.

Although relatively new in the stochastic simulation community, the DPM model has been extensively studied and widely applied in the statistics and machine learning communities during the past decade; see Ferguson (1973), Lo (1984), Escobar (1994), Ghosh and Ramamoorthi (2003), etc. In general, DPM has demonstrated robust performance in terms of density estimation (Escobar and West 1995, Görür and Rasmussen 2010, etc.). The Markov chain Monte Carlo (MCMC) method enables efficient sampling of mixture distributions from the posterior; see for example Escobar and West (1995), Neal (2000), Hanson (2006), Kottas (2006), and Wang and Dunson (2011).

This paper completes and extends our prior work (Xie et al. 2014) on the input uncertainty and the uncertainty quantification for the system performance estimation. Compared with the parametric Bayesian approaches, our nonparametric input models can capture the important features in the real-world data. The posteriors of flexible input models can automatically account for

both model selection and parameter values uncertainty. Compared with BMA, our approach can avoid the difficulty in selecting the "appropriate" parametric candidate distributions. Further, our Bayesian framework accounts for both input and simulation uncertainty in the Bayesian paradigm and delivers a CrI of the system true mean performance instead focusing on estimating the system posterior mean response.

3. A Bayesian Nonparametric Hierarchical Framework

When we use the simulation to assess the stochastic system performance, the output from the jth replication with input models, denoted by F, can be written as

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

where $\mu(F)$ denotes the mean system response and $\epsilon_j(F)$ represents the simulation error following the normal distribution $\epsilon_j(F) \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma_\epsilon^2(F))$. The normality assumption on the simulation error could hold in many situations since the simulation output is often an average of a large number of more basic outputs. For example, in a RM inventory control, when we assess the expected stock level for an ordering policy, each simulation output is the average of stock levels collected from many time periods. This normal assumption does not hold in general. The empirical study on an $M/M/1$ queue with different utilization demonstrates the robustness of our approach in Section 4. For notational simplification, suppose that F consists of a single univariate model. Denote the unknown underlying true input model by F^c.

Our interest is in the system mean response at the true input model, denoted by $\mu^c \equiv \mu(F^c)$. Since the simulation output depends on the choice of input distribution F, the input model failing to capture important features of F^c can lead to poor estimates of system performance. Thus, it is desirable to construct the input model that can faithfully capture the heterogeneity, multi-modality, skewness and tails in the real-world data. Without strong prior information on F^c, in Section 3.1, we present nonparametric DPM that provides a natural way to model the input generative processes with various latent sources of uncertainty. Thus, it can capture the important features in the real-world data.

The underlying true input distribution F^c is estimated by finite real-world data of size m, denoted by $\mathbf{X}_m \equiv \{X_1, X_2, \ldots, X_m\}$, with $X_i \overset{i.i.d.}{\sim} F^c$. The posterior distribution of the *flexible* input model derived from the Bayes' rule can be used to quantify the input uncertainty,

$$p(F|\mathbf{X}_m) \propto p(F) \cdot p(\mathbf{X}_m|F)$$

where $p(F)$ characterizes our prior belief about the true input model F^c and $p(\mathbf{X}_m|F)$ denotes the likelihood function of data $\mathbf{X}_m$ under a generic input model F. Since the DPM model does not have closed form distributions for analytical posterior analysis, we describe Gibbs samplers in Section 3.2 to efficiently draw posterior samples of input models, $\{\tilde{F}^{(1)}, \tilde{F}^{(2)}, \ldots, \tilde{F}^{(B)}\}$, quantifying the input uncertainty. Then, we discuss the asymptotic consistency of $p(F|\mathbf{X}_m)$ in Section 3.3.

When the Bayesian nonparametric approach is used to quantify the input uncertainty, the number of *active* parameters varies at different posterior samples of input model; see the explanation in Sections 3.1 and 3.2. This poses difficulties in constructing an appropriate metamodel as a functional of nonparametric input model. Thus, the direct simulation is used to propagate the input uncertainty to the output. Given finite simulation resource, we characterize the simulation estimation uncertainty by the Bayesian posteriors of mean system responses in Section 3.4. Specifically, at any sample $\tilde{F}^{(b)}$ drawn from $p(F|\mathbf{X}_m)$, we generate n_b replications and obtain the outputs $\mathbf{Y}_b = \{Y_1(\tilde{F}^{(b)}), Y_2(\tilde{F}^{(b)}), \ldots, Y_{n_b}(\tilde{F}^{(b)})\}$. The simulation uncertainty of mean response at $\tilde{F}^{(b)}$ is characterized by the posterior distribution, denoted by $p(\mu_b|\mathbf{Y}_b, \tilde{F}^{(b)})$, where $\mu_b \equiv \mu(\tilde{F}^{(b)})$. Let $\tilde{\mu}_b \equiv \tilde{\mu}(\tilde{F}^{(b)}) \sim p(\mu_b|\mathbf{Y}_b, \tilde{F}^{(b)})$ be a random draw from the posterior.

Thus, our belief on μ^c is characterized by the posterior distribution of the compound random variable $U \equiv \tilde{\mu}(\tilde{F})$, denoted by $F_U(\cdot|\mathbf{X}_m, \mathcal{Y}_\mathbf{n})$, given the information obtained from the real-world data $\mathbf{X}_m$ and the simulation outputs $\mathcal{Y}_\mathbf{n} \equiv \{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_B\}$, where the number of replications allocated to the samples of input distribution $\{\tilde{F}^{(1)}, \tilde{F}^{(2)}, \ldots, \tilde{F}^{(B)}\}$ is collected in the vector $\mathbf{n} \equiv \{n_1, n_2, \ldots, n_B\}$. In Section 3.5.1, we propose a sampling procedure to construct a $(1-\alpha^*)100\%$ *percentile CrI* quantifying the overall estimation uncertainty of μ^c, denoted by $[q_{\alpha^*/2}(\mathbf{X}_m, \mathcal{Y}_\mathbf{n}), q_{1-\alpha^*/2}(\mathbf{X}_m, \mathcal{Y}_\mathbf{n})]$, where

$$q_\gamma(\mathbf{X}_m, \mathcal{Y}_\mathbf{n}) \equiv \inf\{q : F_U(q|\mathbf{X}_m, \mathcal{Y}_\mathbf{n}) \geq \gamma\}$$

with $\gamma = \alpha^*/2, 1-\alpha^*/2$. Then, we study the asymptotic properties of the CrI in Section 3.5.2. As real-world systems have to evolve rapidly to be competitive, the decision makers often face the situations where the amount of real-world data is limited. Thus, we prove that given a finite amount of real-world data $\mathbf{X}_m$, as the simulation budget increases, the interval $[q_{\alpha^*/2}(\mathbf{X}_m, \mathcal{Y}_\mathbf{n}), q_{1-\alpha^*/2}(\mathbf{X}_m, \mathcal{Y}_\mathbf{n})]$ converges to the CrI quantifying the impact of input uncertainty, denoted by $[q_{\alpha^*/2}(\mathbf{X}_m, \mu(\cdot)), q_{1-\alpha^*/2}(\mathbf{X}_m, \mu(\cdot))]$, where

$$q_\gamma(\mathbf{X}_m, \mu(\cdot)) \equiv \inf\{q : F_U(q|\mathbf{X}_m, \mu(\cdot)) \geq \gamma\}$$

and $F_U(\cdot|\mathbf{X}_m, \mu(\cdot))$ denotes the posterior distribution of $\mu(\tilde{F})$ with $\tilde{F} \sim p(F|\mathbf{X}_m)$. We also show that as the real-world data and the simulation budget go to infinity, this interval converges to μ^c.

If the interval $[q_{\alpha^*/2}(\mathbf{X}_m, \mathcal{Y}_\mathbf{n}), q_{1-\alpha^*/2}(\mathbf{X}_m, \mathcal{Y}_\mathbf{n})]$ accounting for both input and simulation uncertainty is too large, the decision maker needs to know if the additional simulation could improve the estimation accuracy of μ^c. Here, we consider the situations where the additional real-world data are not easy to collect. Otherwise, the input uncertainty may not be a concern. Thus, we derive a variance decomposition to estimate the relative contributions from input and simulation uncertainty in Section 3.6.

3.1. Input Modeling by Dirichlet Process Mixtures

Given a kernel density function $h(\cdot)$, an input distribution from DPM is represented as an infinite mixture with the density

$$f(x) = \sum_{j=1}^{+\infty} \pi_j h(x|\boldsymbol{\psi}_j) \tag{1}$$

where π_j denotes the mixing weights and $h(\cdot|\boldsymbol{\psi}_j)$ represents the kernel density function with parameters $\boldsymbol{\psi}_j$. The mixing distribution of parameters $\left\{(\pi_j, \boldsymbol{\psi}_j)_{j=1}^{+\infty}\right\}$, which is $\sum_{j=1}^{+\infty} \pi_j \delta(\boldsymbol{\psi}_j)$ ($\delta(a)$ is the Dirac function at a), is drawn from the Dirichlet process $\mathrm{DP}(\alpha, G_0)$, where G_0 denotes the base distribution and α denotes the dispersion parameter. The weight sequence $\{\pi_1, \pi_2, \ldots\}$ is controlled by α, $\pi_j = \beta_j \Pi_{\ell=1}^{j-1}(1-\beta_\ell)$, with $\beta_j \sim \mathrm{Beta}(1, \alpha)$, and the parameters are generated by G_0, $\boldsymbol{\psi}_j \sim G_0$.

Theoretically, the number of latent sources of uncertainty could be infinity. However, given finite real-world data $\mathbf{X}_m = (X_1, X_2, \ldots, X_m)$, there is a finite number of *active* components, denoted by K. It is bounded by m since each active component requires more than one data points associated with. Let $\mathbf{c} = (c_1, c_2, \ldots, c_m)$ denote the latent indicator variables that give the indices of components that the data $\mathbf{X}_m$ are associated to. Then, DPM model can be written as (Neal 2000),

$$\begin{aligned} \boldsymbol{\pi} &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \\ c_i|\boldsymbol{\pi} &\overset{i.i.d.}{\sim} \text{Multinomial}(\pi_1, \ldots, \pi_K) \\ \boldsymbol{\psi}_j &\overset{i.i.d.}{\sim} G_0(\boldsymbol{\psi}|\boldsymbol{\theta}_G) \\ X_i|c_i, \Psi &\sim h(\cdot|\boldsymbol{\psi}_{c_i}) \end{aligned} \tag{2}$$

for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, K$, where $\boldsymbol{\pi} \equiv (\pi_1, \pi_2, \ldots, \pi_K)$, $\Psi \equiv (\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_K)$ and $\boldsymbol{\theta}_G$ denotes the hyper-parameters for $G_0(\cdot)$. By integrating out $\boldsymbol{\pi}$, the conditional prior of c_i is

$$\text{P}\big(c_i = j|\mathbf{c}^{-i}\big) = \begin{cases} \frac{m_j^{-i}}{m+\alpha-1} \text{ if } \exists c_q = j \text{ for all } q \neq i \\ \frac{\alpha}{m+\alpha-1} \text{ otherwise} \end{cases} \tag{3}$$

where $\mathbf{c}^{-i}$ are all the latent variables except for c_i, and m_j^{-i} is the number of latent variables with $c_q = j$ for all $q \neq i$; see Neal (2000).

DPM is specified by three key components: the dispersion parameter α, the kernel density $h(\cdot)$, and the base distribution G_0. The parameter α represents the complexity of input model and the dispersion of the data. From Equation (3), conditional on all other samples, the probability of assigning X_i to a new mixing component is $\alpha/(m + \alpha - 1)$. DPM with a larger value of α tends to generate samples of the input density $f(\cdot)$ with more distinct active components, which implies higher complexity. The appropriate value of α could be inferred from the real-world data; see the inference in Section 3.2. Therefore, differing from parametric approaches, densities generated as DPM can adapt its complexity to the data.

The choice of the kernel density $h(\cdot)$ is based on the support of F^c, and meanwhile accounts for the feasibility of implementation in the posterior computation. Here, suppose that the support

of underlying input model is known. It could be a limitation for some cases where the support is unknown; see for example Section 6.8 of Law (2015). We present DPM models with three kernel densities, including Gamma, Gaussian and Beta, which account for the real-world data that are supported on the half real line $\Re^+$, the whole real line $\Re$ and a finite interval $[a_1, a_2]$ with $-\infty < a_1 < a_2 < \infty$. Notice that the scaled version of DPM with Beta kernel could be applicable to continuous distributions with a finite support interval. As a result, the DPMs with these kernels can be widely applied to different types of real-world data. Further, since Gamma, Gaussian and Beta distributions belong to the exponential family, with conjugate priors, we derive efficient samplers to generate posterior samples of input model quantifying the input uncertainty.

To simplify the posterior sampling, we consider the conjugate prior G_0 for component parameters $\boldsymbol{\psi}$. For DPM with Gamma kernel, we let $\boldsymbol{\psi} = (V, u)^\top$ with V and u denoting the shape and mean parameters. Motivated by the study on Gamma mixture distributions in Wiper (2001), we consider a conditional conjugate prior for V and u,

$$V \sim \text{Exponential}(\theta) \text{ and } u \sim \text{Inv-Gamma}(\gamma, \beta). \tag{4}$$

Equation (4) specifies $G_0(V, u)$ with the hyper-parameters $\boldsymbol{\theta}_G = (\theta, \gamma, \beta)$.

For DPM with Gaussian kernel, we let $\boldsymbol{\psi} = (u, \sigma^2)^\top$ with u and σ^2 denoting the mean and variance of the Gaussian component. Following Gelman et al. (2004), we choose the conjugate prior,

$$u|\sigma^2 \sim \mathcal{N}(u_0, \sigma^2/k_0) \text{ and } \sigma^2/\sigma_0^2 \sim \text{Inv-Gamma}\left(\frac{v_0}{2}, \frac{1}{2}\right). \tag{5}$$

Equation (5) specifies $G_0(u, \sigma^2)$ with hyper-parameters $\boldsymbol{\theta}_G = (u_0, k_0, v_0, \sigma_0)^\top$.

For DPM with Beta kernel, we let $\boldsymbol{\psi} = (\gamma, \beta)^\top$ with γ and β denoting the two shape parameters. Since the Beta distribution belongs to the exponential family, we choose the conjugate prior,

$$\gamma, \beta|\lambda_0, \lambda_1, \lambda_2 \propto \exp\left\{-\lambda_1\gamma_j - \lambda_2\beta - \lambda_0 \log\left[\frac{\Gamma(\gamma)\Gamma(\beta)}{\Gamma(\gamma+\beta)}\right]\right\}. \tag{6}$$

Equation (6) specifies $G_0(\gamma, \beta)$ with the hyper-parameters $\boldsymbol{\theta}_G = (\lambda_0, \lambda_1, \lambda_2)^\top$.

Based on our empirical study, these base distributions demonstrate good performance. Notice that it is possible to impose hyper-priors on both G_0 and α, which adds more model flexibility and

adaptivity to the data. However, this leads to more complex sampling procedure, and further our empirical studies indicate that adding these hyper-priors has insignificant influence on the posterior inference. Therefore, we only consider the posterior analysis on $\Psi, \mathbf{c}, \alpha$.

3.2. Gibbs Sampler for DPM

The DPM model (2) does not have closed form distributions for analytical posterior analysis. Motivated by Neal (2000), we present a Gibbs sampler for the parameters of distinct components Ψ, the indicator variables $\mathbf{c} = (c_1, c_2, \ldots, c_m)$, and the dispersion parameter α. Each iteration in the Gibbs sampler includes three main steps as follows. In Step (1), for each observation X_i in the real-world data $\mathbf{X}_m$, we update its latent indicator c_i conditional on all the other parameters, and then record the number of active distinct components K. In Step (2), for each component, we update its parameters $\boldsymbol{\psi}_j$ given the data associated to this component. In Step (3), we update the dispersion parameter α conditional on the current number of active components.

(1) For $i = 1$ to m, generate c_i from the conditional posterior $p(c_i = j|\mathbf{c}^{-i}, \boldsymbol{\psi}_j, \alpha, X_i)$. Remove empty components and record the number of active distinct components K.

(2) For $j = 1$ to K, generate the kth parameter in $\boldsymbol{\psi}_j$, denoted by ψ_{jk}, from the conditional posterior $p(\psi_{jk}|\boldsymbol{\psi}_j^{-k}, \mathbf{X}^j)$, where $\boldsymbol{\psi}_j^{-k}$ denotes the remaining parameters in $\boldsymbol{\psi}_j$ and $\mathbf{X}^j$ denotes all the data associated to the j component.

(3) Generate α from the posterior $p(\alpha|K)$.

The posterior inference and sampling for the indicator variables $\mathbf{c}$ and component parameters $\boldsymbol{\psi}_j$ for $j = 1, 2, \ldots, K$ in Steps (1) and (2) can be found in the Appendix. When we update the dispersion parameter α in Step (3), its posterior only depends on the number of active distinct components K, i.e. $p(\alpha|K) \sim p(\alpha)p(K|\alpha)$. We impose a prior, $p(\alpha) = \text{Gamma}(a, b)$, on α, with shape $a > 0$ and scale $b > 0$. Thus, the hyper-parameters for α are $\boldsymbol{\theta}_\alpha = (a, b)^\top$. To simplify the sampling procedure for $p(\alpha|K)$, following Escobar and West (1995), we introduce a new random variable η and generate α from $p(\alpha|K)$ by

$$\eta|\alpha, K \sim \text{Beta}(\alpha + 1, m)$$

$$\alpha|\eta, K \sim \tau\text{Gamma}(a + K, b - \log(\eta)) + (1 - \tau)\text{Gamma}(a + K - 1, b - \log(\eta)).$$

where τ is defined by $\tau/(1-\tau) = (a+K-1)/[m(b-\log(\eta))]$. Notice that the approaches proposed to improve the Gibbs sampling efficiency for DPM through a collapse of the state space of the Markov chain in Maceachern (1994, 1998), MacEachern and Muller (2000) could be incorporated into our Bayesian framework.

Thus, this sampling procedure can generate samples $\{\tilde{F}^{(1)}, \tilde{F}^{(2)}, \ldots, \tilde{F}^{(B)}\}$ quantifying the input uncertainty. Given $\mathbf{c}$ from Step (1), we can estimate the weights $\pi_j = \sum_{i=1}^{m} \delta(c_i = j)/m$ for $j = 1, 2, \ldots, K$ and $i = 1, 2, \ldots, m$, where $\delta(\cdot)$ denotes a Delta function. Combining $\mathbf{\Psi}$ from Step (2), we can get a posterior sample of input model, which is a finite mixture with density $\tilde{f}(x) = \sum_{j=1}^{K} \pi_j h(x|\boldsymbol{\psi}_j)$. To further estimate the system mean response, we can generate input variates by

$$c|\boldsymbol{\pi} \sim \text{Multinomial}(\pi_1, \ldots, \pi_K) \text{ and } X|c \sim h(\cdot|\boldsymbol{\psi}_c) \tag{7}$$

to drive the simulation. Notice that the number of active components, K, can vary at different samples of input model.

3.3. Posterior Consistency of Input Models

In the Bayesian paradigm, a very basic requirement is the *posterior consistency* at the true input distribution (Ghosal et al. 1999). This means that as the amount of real-world data increases, the posterior becomes more and more concentrated near F^c with probability approaching 1. The posterior consistency for DPM is studied in the statistics literature, such as Ghosal et al. (1999), Tokdar (2006), Wu and Ghosal (2008), etc. Given the prior distributions in Equations (4) and (5), Theorem 1 summarizes posterior consistency results on DPM with Gamma and Gaussian kernels for input distributions supported on $\Re^+$ and $\Re$.

The posterior consistency in Theorem 1 is stated in the following sense of *weak consistency*. For two generic distributions (and measures) F_1 and F_2 on $\Re$ with the Borel sigma algebra $\mathcal{B}(\Re)$, their *Lévy-Prokhorov (L-P) distance* (Billingsley 1999) is defined by $d_{LP}(F_1, F_2) \equiv \inf\{\eta > 0 \mid F_1(A) \leq F_2(A^\eta) + \eta \text{ and } F_2(A) \leq F_1(A^\eta) + \eta, \text{ for all } A \in \mathcal{B}(\Re)\}$, where $A^\eta \equiv \{a \in \Re \mid \exists b \in A, |a-b| < \eta\}$. The L-P distance, denoted by d_{LP}, is a metric under which the convergence is equivalent

to the weak convergence of measures on $\Re$. If $\tilde{F}_m$ is drawn from the posterior $p(F|\mathbf{X}_m)$ and $\lim_{m\to\infty} \mathrm{P}\{d_{LP}(\tilde{F}_m, F^c) > \epsilon\} = 0$ for any $\epsilon > 0$, then $\tilde{F}_m$ converges in probability to F^c, and we write $\tilde{F}_m \xrightarrow{P} F^c$ as $m \to \infty$. Then, the posterior $p(F|\mathbf{X}_m)$ is defined as *weakly consistent* at F^c.

THEOREM 1. *Let* $\mathbf{X}_m \equiv \{X_1, X_2, \ldots, X_m\}$ *with* $X_i \overset{i.i.d.}{\sim} F^c$ *for* $i = 1, 2, \ldots, m$.

(i) (Wu and Ghosal 2008 Theorem 14) Suppose the DPM with Gamma kernel has the prior specified as Equation (4). Let f^c *be a continuous and bounded density with support on* $\Re^+$ *satisfying the following conditions: (a)* $f^c(x) \in (0, C_f]$ *for some constant* $0 < C_f < \infty$ *for all* x*; (b)* $|\int_0^\infty f^c(x) \log f^c(x) dx| < \infty$*; (c)* $\int_0^\infty f^c(x) \log \frac{f^c(x)}{\phi_\delta(x)} dx < \infty$ *for some* $\delta > 0$*, where* $\phi_\delta(x) = \inf_{[x,x+\delta)} f^c(t)$ *if* $0 < x < 1$ *and* $\phi_\delta(x) = \inf_{(x-\delta,x]} f^c(t)$ *if* $x \geq 1$*; (d) there exists* $\zeta > 0$ *such that* $\int_0^\infty \max(x^{-\zeta-2}, x^{\zeta+2}) f^c(x) dx < \infty$*. Then, the posterior* $p(F|\mathbf{X}_m)$ *from DPM with Gamma kernel is weakly consistent at* F^c*.*

(ii) (Tokdar 2006 Theorem 3.3) Suppose the DPM with Gaussian kernel has the prior specified as Equation (5). Let F^c *(and the density* f^c*) be supported on* $\Re$ *and assume that it satisfies the following conditions: (a)* $\left|\int_{-\infty}^{+\infty} f^c(x) \log f^c(x) dx\right| < +\infty$*; (b) there exists an* $\eta \in (0,1)$*, such that* $\int_{-\infty}^{+\infty} |x|^\eta f^c(x) dx < +\infty$*; (c) there exist constants* $\sigma_0 > 0, c_1 \in (0, \eta), c_2 > c_1, b_1 > 0, b_2 > 0$*, such that for the base measure* $G_0(u, \sigma)$ *and for all large* $x > 0$*:*

$$\max\left\{G_0\left([x - \sigma_0 x^{\eta/2}, +\infty) \times [\sigma_0, +\infty)\right), G_0\left([0, +\infty) \times (x^{1-\eta/2}, +\infty)\right)\right\} \geq b_1 x^{-c_1};$$

$$\max\left\{G_0\left((-\infty, -x + \sigma_0 x^{\eta/2}] \times [\sigma_0, +\infty)\right), G_0\left((-\infty, 0] \times (x^{1-\eta/2}, +\infty)\right)\right\} \geq b_1 x^{-c_1};$$

$$G_0\left((-\infty, x) \times (0, e^{x^\eta - 1/2})\right) > 1 - b_2 x^{-c_2}; \quad G_0\left((x, +\infty) \times (0, e^{x^\eta - 1/2})\right) > 1 - b_2 x^{-c_2}.$$

Then, the posterior $p(F|\mathbf{X}_m)$ *from DPM with Gaussian kernel is weakly consistent at* F^c*.*

Theorem 1 indicates that the posterior from DPM with Gamma and Gaussian kernels can consistently estimate the true input distributions under very weak conditions on the existence of moments and entropy of F^c, as well as the boundedness and continuity of f^c. Essentially, no assumptions on the analytical forms of F^c and f^c are required for the posterior consistency. Distributions with support on $\Re^+$, including Lognormal, Pearson Type V, Johnson S_B, Johnson S_U, log-logistic, Gamma

and Weibull with shape parameter greater than 1, satisfy the conditions in Part (i) of Theorem 1. Distributions with support on $\Re$, including Normal, Logistic, Student's t and Cauchy, satisfy the conditions in Part (ii) of Theorem 1. For Part (ii), Condition (c) on the base measure G_0 are satisfied by the normal-inverse-gamma prior in Equation (5) if we choose appropriate hyperparameters; see the remarks after Theorem 3.3 in Tokdar (2006). Furthermore, Condition (b) includes many heavy tailed distributions, such as Cauchy distribution and Student's t-distribution with at most two degrees of freedom. Therefore, Part (ii) indicates that the posterior of DPM with Gaussian kernel is weakly consistent for all these heavy tailed distributions.

Conditions (a) and (d) in Part (i) of Theorem 1 may seem restrictive, as they have excluded some distributions, such as the gamma, log-logistic, and Weibull distributions with shape parameters less than or equal to 1. However, our empirical study indicates that the performance of DPM with Gamma kernel is robust to the conditions in Part (i).

In Theorem 1, we do not discuss the weakly consistency of DPM with Beta kernel. Existing Bayesian asymptotic results in the literature mainly focus on slightly different versions of Beta mixtures, such as the finite mixtures of Bernstein polynomials (Petrone and Wasserman 2002, Wu and Ghosal 2008), or the finite Beta mixtures in Rousseau (2010). Ghosal et al. (2008) contains partial results on the classes of distributions that can be expressed an infinite mixture of Betas. In general, the posterior consistency for DPM with Beta kernel is still an open problem. However, our empirical study demonstrate its flexibility and adaptiveness to different types of input models.

3.4. Simulation Estimation Uncertainty Quantification

The Gibbs sampling in Section 3.2 can generate samples $\{\tilde{F}^{(1)}, \tilde{F}^{(2)}, \ldots, \tilde{F}^{(B)}\}$ quantifying the input uncertainty. Without any prior information on the true mean response surface $\mu(\cdot)$, we run simulations at each sample of input model to estimate the mean response. Given a finite simulation budget, the system responses are estimated with error. Thus, at each $\tilde{F}^{(b)}$ with $b = 1, 2, \ldots, B$, we develop the posterior of $\mu(\tilde{F}^{(b)})$ to quantify the simulation uncertainty.

Specifically, at any $\tilde{F}^{(b)}$, we run n_b replications and obtain the outputs $\mathbf{Y}_b = \{Y_1(\tilde{F}^{(b)}), Y_2(\tilde{F}^{(b)}), \ldots, Y_{n_b}(\tilde{F}^{(b)})\}$ with $Y_j(\tilde{F}^{(b)}) \overset{i.i.d.}{\sim} \mathcal{N}(\mu_b, \sigma_b^2)$ for $j = 1, 2, \ldots, n_b$, where $\mu_b = \mu(\tilde{F}^{(b)})$ and $\sigma_b^2 = \sigma_\epsilon^2(\tilde{F}^{(b)})$. By the Bayes' rule, the posterior of system mean response at $\tilde{F}^{(b)}$ is

$$p\left(\mu_b|\tilde{F}^{(b)}, \mathbf{Y}_b\right) \propto p(\mu_b) \cdot \prod_{j=1}^{n_b} \frac{1}{\sqrt{2\pi}\sigma_b} \exp\left\{-\frac{\left[Y_j(\tilde{F}^{(b)}) - \mu_b\right]^2}{2\sigma_b^2}\right\}.$$

Since there is no prior information on the system response, we choose the flat prior $p(\mu_b)$ to be $\mathcal{N}(\mu_0, \sigma_0^2)$ with $\mu_0 = 0$ and $1/\sigma_0^2 = c$, where c is a very small value. Thus, by following the similar derivation with Gelman et al. (2004), the posterior becomes $\mathcal{N}(\widehat{\mu}_b, \widehat{\sigma}_b^2)$ with

$$\widehat{\mu}_b = \frac{\mu_0/\sigma_0^2 + n_b\bar{Y}/\sigma_b^2}{1/\sigma_0^2 + n_b/\sigma_b^2} \text{ and } \frac{1}{\widehat{\sigma}_b^2} = \frac{1}{\sigma_0^2} + \frac{n_b}{\sigma_b^2}$$

where $\bar{Y}_b = \sum_{j=1}^{n_b} Y_j(\tilde{F}^{(b)})/n_b$. By letting $c \to 0$, the posterior $p(\mu_b|\tilde{F}^{(b)}, \mathbf{Y}_b)$ becomes $\mathcal{N}(\bar{Y}_b, \sigma_b^2/n_b)$. Since σ_b^2 is unknown, we consider the empirical Bayesian by replacing it with the sample variance, $S_b^2 = \sum_{j=1}^{n_b}[Y_j(\tilde{F}^{(b)}) - \bar{Y}_b]^2/(n_b - 1)$. Thus, the random draw $\tilde{\mu}_b$ from the posterior distribution characterizing our belief on the system mean response at $\tilde{F}^{(b)}$ is

$$\tilde{\mu}_b|\tilde{F}^{(b)}, \mathbf{Y}_b \sim \mathcal{N}\left(\bar{Y}_b, \frac{S_b^2}{n_b}\right). \tag{8}$$

Our empirical study over an $M/M/1$ in Section 4.2 demonstrates that the performance of our approach is robust to the normal assumption on the simulation estimation error and also the use of plug-in empirical Bayesian in Equation (8).

3.5. Quantify the Overall Estimation Uncertainty for μ^c

Without strong prior information on F^c and $\mu(\cdot)$, the Bayesian hierarchical framework is introduced to quantify the overall estimation uncertainty of $\mu^c = \mu(F^c)$. In Section 3.5.1, we provide the sampling procedure to construct the posterior for the compound random variable $U = \tilde{\mu}(\tilde{F})$ characterizing our belief of μ^c and build a percentile CrI quantifying both input and simulation estimation uncertainty. Then, we study the asymptotic properties of the CrI in Section 3.5.2.

3.5.1. Procedure to Construct the CrI The procedure includes the main steps as follows. Based on the support of input model F^c, choose an appropriate kernel density function $h(\cdot)$, and then specify the hyper-parameters for both G_0 and α in Step 1. See Section 4.1 for the values of hyper-parameters used in our empirical study. In Step 2(a), given the real-world data $\mathbf{X}_m$, generates samples from the posterior of input distribution $\tilde{F}^{(b)} \sim p(F|\mathbf{X}_m)$ with $b = 1, 2, \ldots, B$ to account for the input uncertainty as described in Section 3.2. At each $\tilde{F}^{(b)}$, generate input variates with Equation (7), use them to drive the simulations, and obtain simulation outputs $\mathbf{Y}_b$ with n_b replications in Step 2(b). Then, draw samples $\tilde{\mu}_b$ from the posterior $p(\mu_b|\tilde{F}^{(b)}, \mathbf{Y}_b)$ characterizing the simulation estimation uncertainty in Step 2(c). Thus, $\{\tilde{\mu}_1, \tilde{\mu}_2, \ldots, \tilde{\mu}_B\}$ obtained from this hierarchical sampling are the samples of $U = \tilde{\mu}(\tilde{F})$ from the posterior $F_U(\cdot|\mathbf{X}_m, \mathcal{Y}_\mathbf{n})$ given the information from the real-world data $\mathbf{X}_m$ and simulation experiments with outputs $\mathcal{Y}_\mathbf{n} = \{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_B\}$. We further construct a $(1-\alpha^*)100\%$ percentile CrI quantifying the overall uncertainty of system performance estimation in Step 3.

1. Based on the support of F^c, choose an appropriate kernel density function $h(\cdot)$. Then, specify hyper-parameters $\boldsymbol{\theta}_G$ and $\boldsymbol{\theta}_\alpha$ for the base distribution G_0 and the dispersion parameter α.

2. For $b = 1, 2, \ldots, B$

   (a) Given the real-world data $\mathbf{X}_m$, generate the sample of input model $\tilde{F}^{(b)} \sim p(F|\mathbf{X}_m)$ by following the Gibbs sampling procedure described in Section 3.2.

   (b) At $\tilde{F}^{(b)}$, generate input variates by using Equation (7), run simulations with n_b replications, and obtain the outputs $\mathbf{Y}_b$.

   (c) Generate the posterior sample of response, $\tilde{\mu}_b \sim p(\mu_b|\tilde{F}^{(b)}, \mathbf{Y}_b)$, by using Equation (8).

   Next b

3. Report a $(1-\alpha^*)100\%$ two-sided percentile CrI for μ^c

$$\text{CrI} = \left[\tilde{\mu}_{(\lceil(\alpha^*/2)B\rceil)}, \tilde{\mu}_{(\lceil(1-\alpha^*/2)B\rceil)}\right] \tag{9}$$

with the order statistics $\tilde{\mu}_{(1)} \leq \tilde{\mu}_{(2)} \leq \ldots \leq \tilde{\mu}_{(B)}$.

According to Xie et al. (2014), we need B to be at least one thousand to estimate the percentile CrI. Without any prior information about the mean response $\mu(\cdot)$, in this paper, we assign equal replications to all samples of input distribution $\{\tilde{F}^{(1)}, \tilde{F}^{(2)}, \ldots, \tilde{F}^{(B)}\}$. Since each simulation run can be computationally expensive, a sequential design of experiments could efficiently use the computational budget and reduce the impact of simulation estimation uncertainty on the system performance by finding the optimal setting for $(B, n_1, n_2, \ldots, n_B)$ (Yi and Xie 2017).

3.5.2. Asymptotic Properties of the CrI In this section, we study the asymptotic properties of the CrI constructed from our Bayesian framework in Section 3.5.1. In many situations, it could be hard or expensive to collect more real-world data when we make decisions. Therefore, in Theorem 2 part (i), we show that given finite real-world data $\mathbf{X}_m$, as the simulation budget increases, the CrI constructed by our approach, $[\tilde{\mu}_{(\lceil(\alpha^*/2)B\rceil)}, \tilde{\mu}_{(\lceil(1-\alpha^*/2)B\rceil)}]$, converges to the $(1-\alpha^*)100\%$ percentile CrI induced by the input uncertainty with the true mean response surface $\mu(\cdot)$ known, $[q_{\alpha^*/2}(\mathbf{X}_m, \mu(\cdot)), q_{1-\alpha^*/2}(\mathbf{X}_m, \mu(\cdot))]$. Then, in Theorem 2 part (ii), we show that as the amount of real-world data and the simulation budget go to infinity, the interval, $[\tilde{\mu}_{(\lceil(\alpha^*/2)B\rceil)}, \tilde{\mu}_{(\lceil(1-\alpha^*/2)B\rceil)}]$, shrinks to the true mean response μ^c.

In Theorem 2, the convergence between two credible intervals is measured under the *Hausdorff distance*, denoted by $d_H(\cdot,\cdot)$, which is widely used for measuring the distance between two sets. It has a simplified expression when A_1 and A_2 are both closed intervals: If $A_1 = [a_1, b_1]$ and $A_2 = [a_2, b_2]$, then $d_H(A_1, A_2) = \max(|a_1 - a_2|, |b_1 - b_2|)$. In this case, the convergence under Hausdoff distance is the same as the point-wise convergence for the two endpoints of CrIs.

THEOREM 2. *Let* $n_{\min} = \min(\mathbf{n})$ *and* $n_{\min} \geq (\log B)^3 / \min\left(\min_{b=1,\ldots,B-1}\left(\mu_{[b+1]} - \mu_{[b]}\right)^2, 1\right)$, *where* $\mu_{[1]} \leq \mu_{[2]} \leq \ldots \leq \mu_{[B]}$ *denote the order statistics of* $\mu_1, \mu_2, \ldots, \mu_B$. *Suppose the following conditions hold:*

(1) The posterior distribution $F_U(\cdot|\mathbf{X}_m, \mu(\cdot))$ *is continuous with a density on its support, and its support includes an open set and is connected;*

(2) For almost surely all $\tilde{F} \sim p(F|\mathbf{X}_m)$, *there exists a finite constant* $C_\sigma > 0$ *with* $\sigma_\epsilon^2(\tilde{F}) \leq C_\sigma$;

(3) For any $\epsilon > 0$, there exists a finite $\delta > 0$ such that $|\mu(F) - \mu(F^c)| < \epsilon$ if $d_{LP}(F, F^c) < \delta$;

(4) The posterior distribution $p(F|\mathbf{X}_m)$ is weakly consistent at F^c.

Then,

(i) If Conditions (1) and (2) hold, then the CrI in Equation (9) *satisfies*

$$d_H\left(\left[\tilde{\mu}_{(\lceil(\alpha^*/2)B\rceil)}, \tilde{\mu}_{(\lceil(1-\alpha^*/2)B\rceil)}\right], \left[q_{\alpha^*/2}(\mathbf{X}_m, \mu(\cdot)), q_{1-\alpha^*/2}(\mathbf{X}_m, \mu(\cdot))\right]\right) = O_p\left(\frac{\log B}{\sqrt{n_{\min}}}\right) + O_p\left(\frac{1}{\sqrt{B}}\right) \tag{10}$$

where O_p denotes the order under the conditional measure of the simulation outputs $\mathcal{Y}_{\mathbf{n}}$ given $\left\{\tilde{F}^{(1)}, \tilde{F}^{(2)}, \ldots, \tilde{F}^{(B)}\right\}$ and $\mathbf{X}_m$.

(ii) If Conditions (1) - (4) hold, then as $m, B \to \infty$, the CrI in Equation (9) *converges to the true system response μ^c in posterior probability.*

Condition (1) assumes that the posterior of system response $\mu(\tilde{F})$ with $\tilde{F} \sim p(F|\mathbf{X}_m)$ is a continuous distribution with a regular support. This is mainly used as a regularity condition for showing the convergence of the interval $\left[\tilde{\mu}_{(\lceil(\alpha^*/2)B\rceil)}, \tilde{\mu}_{(\lceil(1-\alpha^*/2)B\rceil)}\right]$. Condition (2) requires that the simulation errors have a bounded variance. Condition (3) is about the continuity of the system response $\mu(F)$ with respect to F around F^c in terms of the L-P distance used in the convergence of input model. The similar continuity assumption is commonly used in the literature on input uncertainty and the Gaussian process metamodel when the parametric family of input model is known; see for example Ankenman et al. (2010), Barton et al. (2014) and Xie et al. (2014). Condition (3) generalizes it to the nonparametric situations. Condition (4) is a direct consequence from Theorem 1 which only provides the asymptotic consistency for input models with support on $\Re^+$ and $\Re$.

Given finite real-world data $\mathbf{X}_m$, Part (i) of Theorem 2 shows that as the simulation budget goes to infinity with $B \to \infty$ and $n_{\min} \to \infty$, the CrI obtained by our Bayesian framework in Equation (9) converges to $[q_{\alpha^*/2}(\mathbf{X}_m, \mu(\cdot)), q_{1-\alpha^*/2}(\mathbf{X}_m, \mu(\cdot))]$. For finite B and $n_{\min}$, it further provides a detailed breakdown of the approximation error from the simulation estimation uncertainty. The first error term in (10) comes from the finite replications ($n_{\min}$) allocated to the posterior samples of input model quantifying the input uncertainty. The $\log B$ term comes from a technical union bound over

all B samples and we can choose $n_{\min}$ sufficiently large to make the first error small. The second error term in (10) comes from using finite (B) posterior samples. The convergence of CrI in (10) is stated in the Bayesian setup conditional on the real-world data $\mathbf{X}_m$ and it does *not* require the sample size $m \to \infty$. Therefore, the bound in Part (i) is *non-asymptotic* in m and only asymptotic in the simulation budget $(n_{\min}, B)$. Part (i) is important and also practically useful since we often face the situations with a limited amount of real-world data. Notice that Part (i) only requires Conditions (1) and (2). Part (ii) is a direct consequence of Part (i) and the posterior consistency of $\tilde{\mu}(\tilde{F})$ at μ^c from Conditions (3) and (4). The detailed proof of Theorem 2 is provided in Appendix B.

3.6. Variance Decomposition

Given the real-world data $\mathbf{X}_m$ and the simulation outputs $\mathcal{Y}_{\mathbf{n}}$, the hierarchical sampling procedure described in Section 3.5 generates samples from the posterior of $U = \tilde{\mu}(\tilde{F})$ accounting for both input and simulation uncertainty. In this section, we develop a variance decomposition to measure the relative contributions from both sources of uncertainty. It provides a guidance on how to reduce the system performance estimation uncertainty if the overall uncertainty of U is too large.

THEOREM 3. *For every fixed $b = 1, 2, \ldots, B$, the total variance of $\tilde{\mu}(F^{(b)})$ with $\tilde{F}^{(b)} \sim p(F|\mathbf{X}_m)$ can be decomposed as*

$$Var\left[\tilde{\mu}(\tilde{F}^{(b)})\middle|\mathbf{X}_m\right] = \sigma_I^2 + \sigma_S^2 \tag{11}$$

where $\sigma_I^2 \equiv Var[\mu_b|\mathbf{X}_m]$ and $\sigma_S^2 \equiv \mathrm{E}\left[2\sigma_b^2/n_b|\mathbf{X}_m\right]$ measure the impacts from the input and simulation uncertainty.

Theorem 3 provides a variance decomposition to quantify the relative contributions from input and simulation uncertainty. The variance component $\sigma_I^2 \equiv \mathrm{Var}[\mu_b|\mathbf{X}_m]$ measuring the impact from input uncertainty decreases as the amount of real-world data increases. For input models with support on $\Re^+$ and $\Re$, as $m \to \infty$, the posterior of input model $p(F|\mathbf{X}_m)$ converges to F^c by Theorem 1, and the impact of input uncertainty disappears $\sigma_I^2 \to 0$ if $\mu(\cdot)$ is continuous around F^c in terms of L-P distance. The variance component $\sigma_S^2 \equiv \mathrm{E}\left[2\sigma_b^2/n_b|\mathbf{X}_m\right]$ measures the impact from

the simulation uncertainty and it is the expected simulation estimation uncertainty weighted by $p(F|\mathbf{X}_m)$. Suppose that $\mu(\cdot)$ and $\sigma_\epsilon^2(\cdot)$ are bounded in the design space. Let $n_{\min} = \min_{b=1,2,\ldots,B} n_b$. Then, as $n_{\min} \to \infty$, the impact of the simulation estimation uncertainty disappears $\sigma_S^2 \to 0$. The detailed derivation of Theorem 3 is provided in Appendix C. Since B is recommended to be at least one thousand, we could ignore the finite sampling uncertainty for the variance estimation.

We can estimate the contributions from both input and simulation uncertainty. At each posterior sample $\tilde{F}^{(b)}$ with $b = 1, 2, \ldots, B$, the response sample mean $\bar{Y}_b$ and variance S_b^2 are asymptotically consistent estimators of μ_b and σ_b^2. Thus, we can estimate the variance components σ_I^2 and σ_S^2 by

$$\widehat{\sigma}_I^2 = \frac{1}{B-1}\sum_{b=1}^{B}\left(\bar{Y}_b - \bar{\bar{Y}}\right)^2 \text{ and } \widehat{\sigma}_S^2 = \frac{2}{B}\sum_{b=1}^{B}\frac{S_b^2}{n_b}, \text{ where } \bar{\bar{Y}} = \frac{1}{B}\sum_{b=1}^{B}\bar{Y}_b. \tag{12}$$

If the width of the CrI in Equation (9) is larger than the desired level, the ratio $\widehat{\sigma}_I/\widehat{\sigma}_S$ can be used to locate the main source of uncertainty and further guide the decision maker on improving the system performance estimation.

4. Empirical Study

We first study the finite-sample performance of nonparametric input models by using simulated data in Section 4.1 and real RM demand data collected from the bio-pharmaceutical manufacturing in Appendix E. Results demonstrate that DPM with appropriate kernel can capture the important properties in real-world data, and it has better and more robust finite-sample performance than existing approaches, including finite mixture, empirical distribution, KDE and parametric distributions. Since some test examples in Section 4.1 violate the conditions in Theorem 1, results also indicate that the performance of DPM is robust to the conditions required for input asymptotic consistency. Then, we use an $M/M/1$ queue and an inventory example with compound Poisson demand to study the performance of our Bayesian framework in Sections 4.2 and Appendix F. Results show that our approach has good and robust performance when there is no strong prior information on the input model and the mean response surface. As the amount of real-world data and the simulation budget increase, the CrI $\left[\tilde{\mu}_{(\lceil(\alpha^*/2)B\rceil)}, \tilde{\mu}_{(\lceil(1-\alpha^*/2)B\rceil)}\right]$ shrinks closer to μ^c. Given

finite real-world data $\mathbf{X}_m$, the probability content (PC) of $\mu(\tilde{F})$ with $\tilde{F} \sim p(F|\mathbf{X}_m)$ located in the CrI is close to the nominal significant level $(1-\alpha^*)100\%$. Further, the ratio $\widehat{\sigma}_I/\widehat{\sigma}_S$ provides a good indicator of the relative contributions from both input and simulation uncertainty. In addition, results of the $M/M/1$ queue under different utilization demonstrate that our approach is robust to the violation of the normal assumption on the simulation estimation error and also the plug-in empirical Bayesian approach used for quantifying the simulation estimation uncertainty.

4.1. Input Density Estimation

In the empirical study, a Gamma prior is used for the dispersion parameter α. Escobar and West (1995) recommend to choose α around 1. We use the prior, $\alpha \sim \text{Gamma}(a, b)$ with $a = 1$ and $b = 1$, which puts a fair degree of support at values around $\alpha = 1$. Our sensitivity study in Appendix D indicates that the input model performance is not sensitive to the values of hyper-parameters $\boldsymbol{\theta}_\alpha$.

In the empirical study, we choose the hyper-parameters $\boldsymbol{\theta}_G$ for the base distribution G_0 as noninformative as possible. We set $\theta = 0.01$, $\gamma = 2$ and $\beta = \bar{X}_m$ for DPM with Gamma kernel density, set $\mu_0 = \bar{X}_m$, $v_0 = 1$, $k_0 = 0.01$ and σ_0 equal to the sample standard deviation of real-world data for DPM with Gaussian kernel density, and set $\lambda_0 = 1, \lambda_1 = \lambda_2 = 0.01$ for DPM with Beta kernel density, where $\bar{X}_m = \sum_{i=1}^m X_i/m$.

Since the real-world data could represent the variability caused by various latent sources of uncertainty, different mixture distributions listed in Table 1 are used to study the finite-sample performance of our nonparametric input models. Example 1 is an exponential distribution with the support on $\Re^+$. We use it to test the robustness of DPM when the conditions for posterior consistency in Part (i) of Theorem 1 are violated. Example 2 is a mixture distribution of Lognormal with the support on $\Re^+$, and it is also used to model the demand for each Poisson arrival in the inventory example in Appendix F. Example 3 is a mixture distribution of Gumbel with the support on $\Re$. Both Lognormal and Gumbel mixtures have heavy tails. Example 4 is a mixture of Beta distributions, which has the support on $[0, 1]$.

Table 1 Test examples to study the input distribution estimation

Example 1	Exponential (exp)	exp(1)
Example 2	Lognormal (L)	0.3L(-0.005,0.1)+0.4L(0.378,0.2)+0.3L(0.654,0.3)
Example 3	Gumbel (Gum)	0.3Gum(1.5,0.1)+0.4Gum(2.5,0.3)+0.3Gum(5,0.5)
Example 4	Beta (Be)	0.3Be(10,90)+0.4Be(20,60)+0.3Be(10,10)

For Bayesian approaches, there exist various model selection criteria, including Bayes Factor (Kass and Raftery 1995), Posterior predictive density (Gelman et al. 2004), and Deviance Information Criteria (Spiegelhalter et al. 2002). However, they are not suitable here since we consider both frequentist and Bayesian candidates. As the Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) test statistics are commonly used to study the goodness of fit in the simulation community, we use KS and AD criteria to study the fitting performance of various approaches, including empirical distribution and finite mixture. Specifically, since the underlying true input model F^c for examples listed in Table 1 are known, we replace the hypothesized distribution in these test statistics with F^c to obtain corresponding distance measures. The KS distance records the largest vertical distance between $F^c(\cdot)$ and the distribution estimated by m real-world data, denoted by $\hat{F}_m(\cdot)$, which could be obtained by different approaches, including DPM with various kernel densities, empirical distribution, KDE and parametric approaches, $D_m \equiv \sup_{x\in\Re}(|F^c(x) - \hat{F}_m(x)|)$. The KS distance puts equal weight to all $x \in \Re$. Since it is typically more challenging to estimate the tail behavior compared to the central part, the AD distance places more weight on the tails of F^c, $A_m^2 \equiv m\int_{-\infty}^{\infty} |F^c(x) - \hat{F}_m(x)|^2 w(x) dF^c(x)$, where the weight function is $w(x) = 1/\left(F^c(x)(1-F^c(x))\right)$. Thus, the AD distance can detect the discrepancies at the tails better.

Table 2 records the statistical behaviors of KS and AD distances (D_m and A_m) obtained by DPM with Gamma, Gaussian and Beta kernel densities, finite mixture (Cheng and Currie 2003), empirical distribution, KDE, and parametric distributions selected based on KS and AD criteria when $m = 50, 100, 500$. All results are based on $N = 1000$ macro-replications. In each macro-replication, we first draw m samples, denoted by $\mathbf{X}_m$, from F^c listed in Table 1 to mimic the procedure collecting m "real-world data". Then, various approaches are used to fit the real-world data, and

calculate the KS and AD distances for the fitted distributions. In the table, "parametric (AD)" and "parametric (KS)" refer to the parametric distributions selected based on the AD and KS statistics by using @Risk. KDE is obtained by using the R function, `kde`, and the bandwidth is selected to minimize the mean integrated squared error (Sheather and Jones 1991). For empirical distribution, KDE and parametric distributions selected baed on the AD and KS statistics, we find the fitted distributions and then record the KS and AD distances for these fitted distributions. Differing from these frequentist approaches that provide the point estimates of input distribution, DPM and the finite mixture proposed in Cheng and Currie (2003) are Bayesian approaches. According to Gelman et al. (2004), the posterior predictive distribution, defined by $f(X|\mathbf{X}_m) = \int f(X|F)dP(F|\mathbf{X}_m)$, is recommended for assessing the fit of input model to the real-world data. Thus, the posterior predictive distribution is used to calculate the KS and AD distances. Specifically, we use the Gibbs samplers described in Section 3.2 and Appendix A to generate 100 samples of input model with the warmup equal to 500 and we save the sample for each 10 draws. Then, we aggregate these samples of input distribution to obtain the posterior predictive distribution. In each macro-replication, we obtain the KS and AD distances, denoted by $D_m^{(b)}$ and $A_m^{(b)}$ with $b = 1, 2, \ldots, N$. After that, we record 95% symmetric CIs for both KS and AD distances, denoted by $\bar{D} \pm 1.96 S_D/\sqrt{N}$ and $\bar{A} \pm 1.96 S_A/\sqrt{N}$, where $\bar{D} = \sum_{b=1}^{N} D_m^{(b)}/N$, $\bar{A} = \sum_{b=1}^{N} A_m^{(b)}/N$, $S_D = \left[\sum_{b=1}^{N}(D_m^{(b)} - \bar{D})^2/(N-1)\right]^{1/2}$ and $S_A = \left[\sum_{b=1}^{N}(A_m^{(b)} - \bar{A})^2/(N-1)\right]^{1/2}$. We highlight the smallest KS and AD distances in Table 2.

From Table 2, we observe that as m increases, the KS and AD distances obtained from all approaches decrease. Even though the parametric distributions are selected based on KS and AD tests, DPM with appropriate kernel density typically has smaller KS and AD distances. Notice that DPM with Gamma and Beta kernel densities performs better than DPM with Gaussian kernel, which is the main focus of study in both statistics and machine learning communities. Further, DPM with Gamma kernel fits different input models with support on $\Re^+$ well. Based on the results of AD distance, DPM can provide better estimation on the tail behavior compared with the empirical, KDE and parametric distributions, especially when m is not large. Note that for Example 1 with the exponential distribution, parametric families can fit the data well. However, when the input model becomes complex, parametric distributions cannot fit the data well.

4.2. An $M/M/1$ Queue and An Inventory Example

An $M/M/1$ queue and an inventory example are used to study the performance of our Bayesian framework. We first consider the $M/M/1$ queue. Suppose that the arrival process is known with the arrival rate equal to $\lambda = 0.5, 0.7, 0.9$. The distribution of service time is $\exp(\tau^c)$ with the rate $\tau^c = 1$. Thus, the underlying utilization $\rho^c = \lambda/\tau^c$ is 0.5, 0.7 and 0.9. We are interested in the expected waiting time in the system and the unknown true response is $\mu^c = 1/(\tau^c - \lambda)$.

To evaluate the performance of our approach, we pretend that the underlying distribution for service time is unknown and it is estimated by m observations drawn from F^c. Since the exponential distribution has support $\Re^+$, we use DPM with Gamma kernel to estimate the input distribution. Empirical distribution, KDE and parametric approaches studied in Section 4.1 are frequentist approaches. Since frequentist and Bayesian approaches have different perspectives on quantifying the uncertainty and we also assess their performance differently (Xie et al. 2014), we focus on studying the finite-sample behavior of our Bayesian framework here. Also, suppose that the mean response is unknown and estimated by the simulation. Each simulation run starts with the empty system, and we set both warmup and runlength equal to 1000 customers. *Since the exponential service distribution violates the conditions in Theorem 1 and also the normal assumption on the simulation estimation error does not hold when the utilization ρ^c is high, this example could be used to study the robustness of our approach.*

By following the sampling procedure described in Section 3.5, we generate samples $\tilde{\mu}_b$ from the posterior $F_U(\cdot|\mathbf{X}_m, \mathcal{Y}_n)$ with $b = 1, 2, \ldots, B$, estimate the posterior mean $\mathrm{E}[U|\mathbf{X}_m, \mathcal{Y}_n]$, and construct the 95% percentile CrI, denoted by $\widetilde{\mathrm{CrI}} = [\tilde{\mu}_{(\lceil(\alpha^*/2)B\rceil)}, \tilde{\mu}_{(\lceil(1-\alpha^*/2)B\rceil)}]$, accounting for both input and simulation uncertainty. To evaluate the performance of our approach, we first record the mean and standard deviation (SD) of the deviation of posterior mean from μ^c, defined by $\mathrm{err} = |\mathrm{E}[U|\mathbf{X}_m, \mathcal{Y}_n] - \mu^c|$. Then, we record the mean and SD of the CrI width, denoted by $|\mathrm{CrI}|$. The probability content (PC) of $F_U(\cdot|\mathbf{X}_m, \mu(\cdot))$ located in $\widetilde{\mathrm{CrI}}$,

$$\mathrm{PC}(\widetilde{\mathrm{CrI}}) = \int_{\widetilde{\mathrm{CrI}}} dF_U(q|\mathbf{X}_m, \mu(\cdot)),$$

Table 2 KS and AD distances obtained from DPM with Gamma, Gaussian and Beta kernel densities, empirical distribution, KDE and parametric distributions selected based on KS and AD tests.

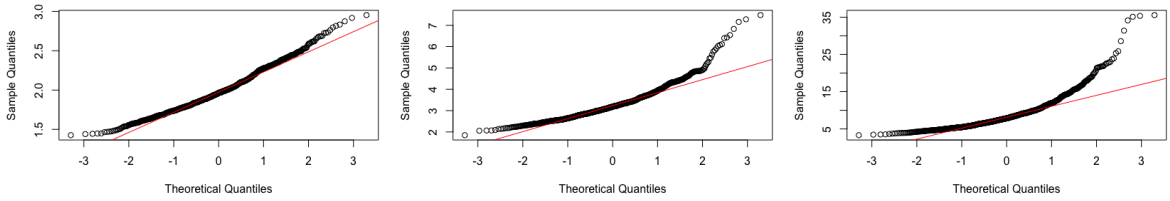
$m = 50$		Example 1	Example 2	Example 3	Example 4
DPM with Gamma	D_m	**0.103±0.002**	**0.071±0.001**	NA	0.072±0.001
	A_m	**11.344±0.147**	**7.336±0.092**	NA	**8.603±0.102**
DPM with Gaussian	D_m	0.109±0.002	0.082±0.001	**0.072±0.001**	0.081±0.001
	A_m	11.641±0.154	9.125±0.117	**6.018±0.097**	9.684±0.120
DPM with Beta	D_m	NA	NA	NA	**0.069±0.001**
	A_m	NA	NA	NA	8.626±0.094
Finite Mixture	D_m	0.147±0.002	0.115±0.002	0.095±0.002	0.085±0.001
	A_m	14.328±0.166	10.824±0.127	9.674±0.115	10.377±0.140
Empirical Distribution	D_m	0.127±0.002	0.099±0.002	0.082±0.001	0.085±0.001
	A_m	13.451±0.162	8.085±0.096	6.679±0.747	10.126±0.139
KDE	D_m	0.112±0.002	0.076±0.001	0.124±0.002	0.089±0.001
	A_m	12.406±0.141	8.429±0.098	10.674±0.132	10.572±0.138
Parametric (KS)	D_m	0.104±0.001	0.080±0.001	0.125±0.002	0.109±0.002
	A_m	11.709±0.136	8.825±0.083	12.107±0.138	11.243±0.148
Parametric (AD)	D_m	0.105±0.002	0.080±0.001	0.125±0.002	0.112±0.002
	A_m	11.684±0.134	8.245±0.087	11.937±0.001	11.095±0.146
$m = 100$		Example 1	Example 2	Example 3	Example 4
DPM with Gamma	D_m	0.085±0.001	**0.054±0.001**	NA	0.050±0.001
	A_m	**8.406±0.107**	**5.283±0.102**	NA	**6.421±0.102**
DPM with Gaussian	D_m	0.088±0.001	0.065±0.001	**0.058±0.001**	0.054±0.001
	A_m	9.278±0.133	5.930±0.098	**4.782±0.061**	7.208±0.117
DPM with Beta	D_m	NA	NA	NA	**0.048±0.001**
	A_m	NA	NA	NA	6.576±0.081
Finite Mixture	D_m	0.107±0.002	0.064±0.001	0.063±0.001	0.059±0.001
	A_m	11.328±0.149	5.824±0.087	5.374±0.093	7.582±0.096
Empirical Distribution	D_m	0.114±0.002	0.071±0.001	0.064±0.001	0.060±0.001
	A_m	12.370±0.155	6.132±0.098	5.016±0.089	7.467±0.093
KDE	D_m	0.109±0.002	0.060±0.001	0.081±0.001	0.058±0.001
	A_m	11.477±0.151	5.709±0.081	6.258±0.094	8.295±0.101
Parametric (KS)	D_m	**0.084±0.001**	0.076±0.001	0.087±0.002	0.084±0.001
	A_m	8.735±0.094	7.645±0.076	8.025±0.118	9.328±0.130
Parametric (AD)	D_m	0.085±0.001	0.077±0.001	0.088±0.002	0.084±0.002
	A_m	8.409±0.091	7.267±0.071	7.742±0.106	9.267±0.126

$m = 500$		Example 1	Example 2	Example 3	Example 4
DPM with Gamma	D_m	**0.061±0.001**	**0.028±0.001**	NA	0.028±0.001
	A_m	5.723±0.079	**2.388±0.052**	NA	3.947±0.058
DPM with Gaussian	D_m	0.064±0.001	0.029±0.001	**0.027±0.001**	0.032±0.001
	A_m	6.038±0.133	2.532±0.048	**2.618±0.047**	4.175±0.069
DPM with Beta	D_m	NA	NA	NA	**0.025±0.001**
	A_m	NA	NA	NA	**3.243±0.042**
Finite Mixture	D_m	0.076±0.001	0.032±0.001	0.039±0.001	0.033±0.001
	A_m	8.314±0.112	2.731±0.051	4.064±0.061	5.096±0.083
Empirical Distribution	D_m	0.078±0.001	0.035±0.001	0.036±0.001	0.036±0.001
	A_m	8.962±0.117	2.874±0.050	3.727±0.054	4.529±0.066
KDE	D_m	0.083±0.001	0.030±0.001	0.045±0.001	0.031±0.001
	A_m	9.633±0.124	3.097±0.054	5.880±0.089	4.827±0.070
Parametric (KS)	D_m	**0.061±0.001**	0.069±0.001	0.074±0.001	0.064±0.001
	A_m	5.875±0.072	6.782±0.038	7.341±0.062	7.343±0.055
Parametric (AD)	D_m	0.062±0.001	0.070±0.001	0.076±0.001	0.067±0.001
	A_m	**5.606±0.071**	6.462±0.031	7.208±0.067	6.905±0.049

is used to evaluate the CrI constructed through our approach. To estimate $\mathrm{PC}(\widetilde{\mathrm{CrI}})$, we draw $B = 1000$ posterior samples of the input models $\tilde{F}^{(b)} \sim p(F|\mathbf{X}_m)$ with $b = 1, \ldots, B$. At each $\tilde{F}^{(b)}$, the expected time staying in the system is $\mu(\tilde{F}^{(b)}) = (1 + \frac{1+C_v^2}{2}\frac{\rho^{(b)}}{1-\rho^{(b)}})M_1^{(b)}$, where $M_1^{(b)}$ and $M_2^{(b)}$ denote the first and second moments of $\tilde{F}^{(b)}$, $\rho^{(b)} = \lambda M_1^{(b)}$ and $C_v^2 = M_2^{(b)}/(M_1^{(b)})^2$. The PC is estimated with $\widehat{\mathrm{PC}}(\widetilde{\mathrm{CrI}}) = \frac{1}{B}\sum_{b=1}^{B} \delta\left(\mu(\tilde{F}^{(b)}) \in \widetilde{\mathrm{CrI}}\right)$. In addition, we calculate the ratio $\widehat{\sigma}_I/\widehat{\sigma}_S$ to estimate the relative contributions from input and simulation uncertainty, where $\widehat{\sigma}_I$ and $\widehat{\sigma}_S$ are obtained by using Equation (12).

When we study the finite-sample performance of our approach, we generate $B = 1000$ posterior samples of input model $\{\tilde{F}^{(1)}, \tilde{F}^{(2)}, \ldots, \tilde{F}^{(B)}\}$ to quantify the input uncertainty and assign equal replications n to each sample. The amount of real-world data m controls the input uncertainty and the number of replications n controls the simulation uncertainty. As the utilization increases, there is more obvious skewness and tail in the simulation outputs. Thus, ρ^c is used to control the violation of the normal assumption on the simulation estimation error; see the Normal Q-Q plots

Figure 1 Normal Q-Q Plots of the Simulation Outputs



(a) When $\rho^c = 0.5$ (b) When $\rho^c = 0.7$ (c) When $\rho^c = 0.9$

for $\rho^c = 0.5, 0.7, 0.9$ in Figure 4.2. The results are from 1000 replications. Even though the normal assumption holds reasonably well when $\rho^c = 0.5$, there exist strong skewness and tail when $\rho^c = 0.9$.

Notice that the proportion of unstable posterior samples of input model $\tilde{F}^{(b)}$ for $b = 1, 2, \ldots, B$, defined as those with the utilization $\rho^{(b)}$ greater and equal to one, increases as m decreases and ρ^c increases. Based on a side experiment, mean and SD (in the bracket) of the percentage of unstable posterior samples estimated by 1000 replications are: 0.009 (0.014) when $m = 50$ and $\rho^c = 0.5$; 0.011 (0.008) when $m = 50$ and $\rho^c = 0.7$; 0.285 (0.0082), 0.17 (0.055) and 0.062 (0.019) when $m = 50, 100, 500$ and $\rho^c = 0.9$. The unstable issue is negligible when $m = 100, 500$ and $\rho^c = 0.5, 0.7$. For simplification, we set the mean response at unstable samples of input model to be infinity. When $\rho^c = 0.9$, the percentage of unstable posterior samples is greater than $\alpha^*/2 = 2.5\%$. Thus, we record the one-sided percentile CrI, denoted by $[\tilde{\mu}_{(\lceil \alpha^* B \rceil)}, +\infty)$.

The results with $m = 50, 100, 500$, $n = 100, 1000$ and $\rho^c = 0.5, 0.7, 0.9$ are shown in Table 3. They are estimated based on 100 macro-replications. From Table 3, as m or n increases, the posterior mean of system mean response becomes closer to μ^c, and the overall estimation uncertainty of $\tilde{\mu}$ gets smaller. The PC is close to the nominal value 95%, which indicates that our approach is robust to the normal assumption on the simulation estimation error and the plug-in empirical Bayesian approach used for quantifying the simulation estimation uncertainty. In addition, the ratio $\widehat{\sigma}_I/\widehat{\sigma}_S$ provides a good measure of the relative contributions from the input and simulation uncertainty.

In addition, we use a RM inventory example to study the performance of the proposed Bayesian framework. The arrivals of move order follow a Poisson process with rate equal to 3. The accumulated move order in the ith time period is $X_i = \sum_{k=1}^{N_i} D_k$, where N_i denotes the number of

Table 3 Results of the $M/M/1$ queue when $m = 50, 100, 500$, $n = 100, 1000$ and $\rho^c = 0.5, 0.7, 0.9$.

		Mean and SD of $\lvert\text{CrI}\rvert/2$	Mean and SD of $\lvert\text{err}\rvert$	Mean and SD of $\widehat{\text{PC}}(\widetilde{\text{CrI}})$	$\widehat{\sigma}_I/\widehat{\sigma}_S$
$\rho^c = 0.5$	$m = 50, n = 100$	1.098 (0.140)	0.652 (0.058)	94.7% (0.7%)	1.779
	$m = 50, n = 1000$	0.925 (0.125)	0.597 (0.050)	94.5% (0.8%)	9.046
	$m = 100, n = 100$	0.847 (0.109)	0.448 (0.034)	94.4% (0.8%)	0.855
	$m = 100, n = 1000$	0.762 (0.086)	0.370 (0.027)	94.3% (0.9%)	4.890
	$m = 500, n = 100$	0.603 (0.081)	0.315 (0.021)	94.1% (0.9%)	0.562
	$m = 500, n = 1000$	0.525 (0.072)	0.269 (0.016)	94.0%(1.0%)	2.527
		Mean and SD of $\lvert\text{CrI}\rvert/2$	Mean and SD of $\lvert\text{err}\rvert$	Mean and SD of $\widehat{\text{PC}}(\widetilde{\text{CrI}})$	$\widehat{\sigma}_I/\widehat{\sigma}_S$
$\rho^c = 0.7$	$m = 50, n = 100$	1.706 (0.224)	0.886 (0.077)	94.6% (0.8%)	1.462
	$m = 50, n = 1000$	1.474 (0.189)	0.802 (0.072)	94.6% (0.7%)	8.389
	$m = 100, n = 100$	1.355 (0.162)	0.723 (0.065)	94.5% (0.9%)	0.767
	$m = 100, n = 1000$	1.149 (0.143)	0.659 (0.054)	94.3% (0.9%)	4.262
	$m = 500, n = 100$	0.958 (0.115)	0.588 (0.048)	94.2% (0.8%)	0.533
	$m = 500, n = 1000$	0.820 (0.096)	0.515 (0.044)	94.2%(0.9%)	2.074
		Mean and SD of $\tilde{\mu}_{(\lceil \alpha^* B \rceil)}$	Mean and SD of $\lvert\text{err}\rvert$	Mean and SD of $\widehat{\text{PC}}(\widetilde{\text{CrI}})$	$\widehat{\sigma}_I/\widehat{\sigma}_S$
$\rho^c = 0.9$	$m = 50, n = 100$	6.093 (0.739)	4.745 (0.586)	94.9% (0.4%)	1.142
	$m = 50, n = 1000$	6.428 (0.784)	4.283 (0.521)	94.8% (0.5%)	7.861
	$m = 100, n = 100$	7.304 (0.832)	3.308 (0.427)	94.7% (0.7%)	0.670
	$m = 100, n = 1000$	7.686 (0.850)	2.794 (0.360)	94.7% (0.6%)	3.283
	$m = 500, n = 100$	8.053 (0.884)	2.591 (0.318)	94.5% (0.8%)	0.452
	$m = 500, n = 1000$	8.244 (0.903)	2.042 (0.266)	94.6%(0.8%)	1.749

arrivals occurring in the ith time period and the size of each move order, denoted by D_k, follows the Log-normal mixture distribution $0.3L(-0.005, 0.1) + 0.4L(0.378, 0.2) + 0.3L(0.654, 0.3)$. DPM with Gamma kernel is used to model the underlying unknown distribution of accumulated move order in each time period. We are interested in the steady-state expected inventory level, type-I and type-II service levels. The empirical results recorded in Appendix F also demonstrate the good performance of our approach.

5. Conclusions

Without strong prior information on the true input models and the system mean response surface, in this paper, a Bayesian nonparametric hierarchical framework is proposed to quantify the overall uncertainty of system performance estimates. Nonparametric DPM can capture the important properties in the real-world data, including heterogeneity, multi-modality, skewness and tails. The posteriors of flexible input models can automatically account for both model selection and parameters value uncertainty. Then, the direct simulation is used to propagate the input uncertainty to the outputs with the simulation uncertainty quantified by the posteriors of system responses. Therefore, our framework leads to a sampling procedure that can deliver a posterior distribution of the system mean response and provide a percentile CrI accounting for both input and simulation uncertainty. A variance decomposition is further developed to quantify the relative contributions from both sources of uncertainty. Our approach is supported with a rigorous asymptotic study. Given a finite amount of real-world data, as the simulation budget increases, our CrI converges to the CrI accounting for input uncertainty with the true mean response surface known. As both real-world data and simulation budget go to infinity, our CrI converges to the true system response.

An empirical study demonstrates obvious advantages of Bayesian nonparametric approaches for input density estimation compared to existing approaches, including empirical distribution, KDE and parametric approaches. The simulation results for an $M/M/1$ queue demonstrate that our framework is robust to the normal assumption on the simulation estimation error. As the amount of real-world data and the simulation budget increase, the CrI accounting for both input and simulation uncertainty shrinks closer to the true system mean response and the probability content of the CrI on $F_U(\cdot|\mathbf{X}_m, \mu(\cdot))$ is close to the nominal value. The ratio σ_I/σ_S provides a good measure of the relative contributions from both sources of uncertainty.

Appendix A: Gibbs Samplers for DPM with Gamma, Gaussian and Beta Kernels

For DPM with Gamma, Gaussian and Beta kernels, we provide the posterior inference and sampling for the indicator variables $\mathbf{c}$ and component parameters $\boldsymbol{\psi}_j$ for $j = 1, 2, \ldots, K$ used in Steps (1) and (2) of the Gibbs samplers presented in Section 3.2. We describe the main results to support the Gibbs sampling in Section A.1. Then, in Section A.2, we provide the detailed derivation of the results used in the sampling procedure.

A.1. Gibbs Sampling for c and ψ

A.1.1. DPM with Gamma Kernel Here, we present a posterior sampler for the DPM with Gamma kernel. Given the base distribution G_0 in Equation (4), we first generate samples of latent variables $\mathbf{c}$ for Step (1) of the Gibbs sampler in Section 3.2. By the Bayes' rule, the conditional posterior of c_i with $i = 1, 2, \ldots, m$ given other variables is

$$p(c_i = j|\mathbf{c}^{-i}, \boldsymbol{\psi}_j, \alpha, X_i) \propto p(c_i = j|\mathbf{c}^{-i}, \boldsymbol{\psi}_j, \alpha) \cdot p(X_i|\mathbf{c}^{-i}, \boldsymbol{\psi}_j, \alpha, c_i = j)$$

where the prior $p(c_i = j|\mathbf{c}^{-i}, \boldsymbol{\psi}_j, \alpha)$ is obtained from Equation (3). Then, the conditional posterior probabilities of c_i in which X_i is associated with either an existing component or a new component with parameters $\boldsymbol{\psi} = (V, u)$ drawn from the base distribution G_0 are

$$p(c_i = j|\mathbf{c}^{-i}, \boldsymbol{\psi}_j, \alpha, X_i) = \begin{cases} b_0 \frac{m_j^{-i}}{m+\alpha-1} X_i^{V_j - 1} e^{-\frac{V_j}{u_j} X_i} \text{ if } \exists c_q = j \text{ for all } q \neq i \\ b_0 \frac{\alpha}{m+\alpha-1} \int X_i^{V-1} e^{-\frac{V}{u} X_i} dG_0(V, u) \text{ otherwise} \end{cases} \tag{13}$$

where b_0 denotes the normalizing constant. When X_i comes from a new component, the conditional posterior for c_i in Equation (13) is not analytically tractable and a sampling approach is used to generate samples of $\mathbf{c}$ by following Algorithm 4 in Neal (2000).

Next we generate samples of the parameters $\boldsymbol{\psi}_j = (V_j, u_j)^\top$ for Step (2) of the Gibbs sampler. By the Bayes' rule, $p(V_j|u_j, \mathbf{X}^j) \propto p(V_j) f(\mathbf{X}^j|V_j, u_j)$ and $p(u_j|V_j, \mathbf{X}^j) \propto p(u_j) f(\mathbf{X}^j|V_j, u_j)$, the conditional posteriors of V_j and u_j are given by

$$\begin{aligned} V_j|u_j, \mathbf{X}^j &\propto \frac{V_j^{m_j V_j}}{\Gamma(V_j)^{m_j}} \exp\left[-V_j \left(\theta + \frac{\sum_{k=1}^{m_j} X_k^j}{u_j} + m_j \log(u_j) - \sum_{k=1}^{m_j} \log(X_k^j)\right)\right] \\ u_j|V_j, \mathbf{X}^j &\sim \text{Inv-Gamma}\left(\gamma + m_j V_j, \beta + V_j \sum_{k=1}^{m_j} X_k^j\right) \end{aligned} \tag{14}$$

where X_k^j are the kth observation associated to the jth component and m_j is the size of $\mathbf{X}^j$. The detailed derivation for these posteriors can be found in Appendix A.2.1.

The conditional posterior $p(V_j|u_j, \mathbf{X}^j)$ in Equation (14) is not a standard distribution. A Metropolis-Hasting (M-H) nested Gibbs sampler is developed to generate samples of V_j from the conditional posterior. Specifically, denote the sample from the previous iteration in the nested M-H sampling by V_j^0. We first generate a candidate sample $\tilde{V}_j$ from a proposal distribution, denoted by $g(\cdot, V_j^0)$, and accept it with probability

$$\min\left\{1, \frac{p(\tilde{V}_j|u_j, \mathbf{X}^j) g(V_j^0, \tilde{V}_j)}{p(V_j^0|u_j, \mathbf{X}^j) g(\tilde{V}_j, V_j^0)}\right\},$$

where $p(V_j^0|u_j, \mathbf{X}^j)$ and $p(\tilde{V}_j|u_j, \mathbf{X}^j)$ are the conditional posteriors from Equation (14). Otherwise, retain the value of V_j^0. The proposal distribution $g(\cdot, V_j^0)$ is chosen to be Gamma$(r, r/V_j^0)$ with

mean located at V_j^0. This proposal distribution is determined by using the Stirling approximation so that it can capture the tail of the conditional posterior $p(V_j|u_j, \mathbf{X}^j)$ well. The detailed derivation can be found in Appendix A.2.1. To make the proposal distribution relatively flat, we recommend that the value of r is set to be small, e.g., $r = 2$ used in our empirical study.

A.1.2. DPM with Gaussian Kernel Given the base distribution G_0 in Equation (5), we first generate samples of the latent variables $\mathbf{c}$ for Step (1) of the Gibbs sampler. If c_i is associated with an existing jth component, then

$$p(c_i = j|\mathbf{c}^{-i}, \boldsymbol{\psi}_j, \alpha, X_i) = b_0 \frac{m_j^{-i}}{m + \alpha - 1} \frac{1}{\sqrt{2\pi}\sigma_j} e^{-\left(X_i - u_j\right)^2 / 2\sigma_j^2}.$$

If c_i is associated with a new component, then

$$p(c_i = j|\mathbf{c}^{-i}, \boldsymbol{\psi}_j, \alpha, X_i) = b_0 \frac{\alpha}{m + \alpha - 1} \frac{(v_0/2)^{v_0/2}}{\Gamma(v_0/2)} \sigma_0^{v_0} \sqrt{\frac{k_0}{2\pi(k_0+1)}} \frac{\Gamma(A)}{B^A}$$

where $A = (v_0 + 1)/2$, $B = [v_0\sigma_0^2 + k_0(X_i - u_0)^2/(k_0 + 1)]/2$ and b_0 is the normalizing constant. The detailed derivation for this conditional posterior can be found in Appendix A.2.2.

Next we generate samples of the parameters $\boldsymbol{\psi}_j = (u_j, \sigma_j^2)^\top$ for Step (2) of the Gibbs sampler. The conditional posteriors for u_j and σ_j are derived by following Chapter 3 in Gelman et al. (2004)

$$\begin{aligned} u_j|\sigma_j, \mathbf{X}^j &\sim \mathcal{N}\left(\frac{k_0}{k_0 + m_j} u_0 + \frac{m_j}{k_0 + m_j} \bar{X}^j, \frac{\sigma_{0j}^2}{k_0 + m_j}\right), \\ \sigma_j^2/\sigma_0^2 \Big| \mathbf{X}^j &\sim \text{Inv-Gamma}\left(\frac{v_0 + m_j}{2}, \frac{1}{2}\right), \end{aligned}$$

where

$$\sigma_{0j}^2 = \frac{v_0\sigma_0^2 + \sum_{k=1}^{m_j}(X_k^j - \bar{X}^j)^2 + \frac{k_0 m_j (\bar{X}^j - u_0)^2}{k_0 + m_j}}{v_0 + m_j} \text{ with } \bar{X}^j = \frac{1}{m_j}\sum_{k=1}^{m_j} X_k^j.$$

A.1.3. DPM with Beta Kernel Here we develop a posterior sampler for DPM with the Beta kernel density to fit the input models with compact supports. We assume that $X_i|c_i = j, \gamma_j, \beta_j \sim \text{Beta}(\gamma_j, \beta_j)$ and denote the parameters for the jth component by $\boldsymbol{\psi}_j = (\gamma_j, \beta_j)^\top$. Equation (6) provides the base function $G_0(\gamma, \beta)$. The derivation for this prior can be founded in Appendix A.2.3.

We first generate samples of the latent variable c_i for Step (1) of the Gibbs sampler. By applying the Bayes' rule, we get the conditional posterior probabilities of c_i, in which X_i is associated with either an existing component or a new component with parameters $\boldsymbol{\psi} = (\gamma, \beta)^\top$ drawn from the base distribution $G_0(\gamma, \beta)$

$$p(c_i = j|\mathbf{c}^{-i}, \boldsymbol{\psi}_j, \alpha, X_i) = \begin{cases} b_0 \frac{m_j^{-i}}{m+\alpha-1} X_i^{\gamma_j - 1}(1 - X_i)^{\beta_j - 1} \text{ if } \exists c_q = j \text{ for all } q \neq i \\ b_0 \frac{\alpha}{m+\alpha-1} \int X_i^{\gamma-1}(1 - X_i)^{\beta-1} dG_0(\gamma, \beta) \text{ otherwise} \end{cases}$$

where b_0 denotes the normalizing constant. Since the conditional posterior for c_i associated with a new component does not have a closed form, we use the sampling approach by following Algorithm 4 in Neal (2000) to generate samples of c_i.

Next we generate samples of the parameters $\boldsymbol{\psi}_j = (\gamma_j, \beta_j)^\top$ for Step (2) of the Gibbs sampler. By applying the Bayes' rule, $p(\gamma_j|\beta_j, \mathbf{X}^j) \propto p(\gamma_j)p(\mathbf{X}^j|\gamma_j, \beta_j)$ and $p(\beta_j|\gamma_j, \mathbf{X}^j) \propto p(\beta_j)p(\mathbf{X}^j|\gamma_j, \beta_j)$, the conditional posteriors of component parameters γ_j and β_j are given by

$$\gamma_j|\beta_j, \mathbf{X}^j \propto \exp\left\{\left[-\lambda_1 + \sum_{k=1}^{m_j}\log(X_k^j)\right]\gamma_j - (\lambda_0 + m_j)\log\left[\frac{\Gamma(\gamma_j)}{\Gamma(\gamma_j+\beta_j)}\right]\right\}, \tag{15}$$

$$\beta_j|\gamma_j, \mathbf{X}^j \propto \exp\left\{\left[-\lambda_2 + \sum_{k=1}^{m_j}\log(1-X_k^j)\right]\beta_j - (\lambda_0 + m_j)\log\left[\frac{\Gamma(\beta_j)}{\Gamma(\gamma_j+\beta_j)}\right]\right\}. \tag{16}$$

The detailed derivation for these posteriors can be found Appendix A.2.3.

Since the conditional posteriors in Equations (15) and (16) are not standard distributions, we again develop an M-H nested Gibbs sampler to generate samples for γ_j and β_j. Denote the samples from the previous iteration in the M-H sampling by γ_j^0 and β_j^0. By using the Stirling approximation, we choose Gamma$(r, r/a)$ with relatively small r and mean a equal to γ_j^0 or β_j^0 as the proposal distribution; See the detailed derivation in Appendix A.2.3. Denote the proposal density by $g(\cdot, a)$. Specifically, for γ_j, we randomly sample a candidate $\tilde{\gamma}_j$ from the proposal distribution Gamma$(r, r/\gamma_j^0)$, and accept $\tilde{\gamma}_j$ with probability

$$\min\left\{1, \frac{p(\tilde{\gamma}_j|\beta_j^0, \mathbf{X}^j)g(\gamma_j^0, \tilde{\gamma}_j)}{p(\gamma_j^0|\beta_j^0, \mathbf{X}^j)g(\tilde{\gamma}_j, \gamma_j^0)}\right\},$$

where $p(\tilde{\gamma}_j|\beta_j^0, \mathbf{X}^j)$ and $p(\gamma_j^0|\beta_j^0, \mathbf{X}^j)$ are the conditional posterior in Equation (15). Otherwise, retain the value of γ_j^0. Similarly, for β_j, we randomly sample a candidate $\tilde{\beta}_j$ from the proposal distribution Gamma$(r, r/\beta_j^0)$, and accept $\tilde{\beta}_j$ with probability

$$\min\left\{1, \frac{p(\tilde{\beta}_j|\gamma_j^0, \mathbf{X}^j)g(\beta_j^0, \tilde{\beta}_j)}{p(\beta_j^0|\gamma_j^0, \mathbf{X}^j)g(\tilde{\beta}_j, \beta_j^0)}\right\},$$

where $p(\tilde{\beta}_j|\gamma_j^0, \mathbf{X}^j)$ and $p(\beta_j^0|\gamma_j^0, \mathbf{X}^j)$ are the conditional posteriors in Equation (16). Otherwise, retain the value of β_j^0. In our empirical study, we set $r = 2$ when we sample both γ_j and β_j.

A.2. Derivation of the Results Used in the Gibbs Sampling

In this section, we provide the detailed derivation of priors, proposal distributions, and conditional posteriors used in the Gibbs samplers for DPM with Gamma, Gaussian and Beta kernel densities in Section A.1.

A.2.1. Conditional Posteriors of DPM with Gamma Kernel We derive the conditional posteriors of parameters $\boldsymbol{\psi}_j = (V_j, u_j)$ with $j = 1, 2, \ldots, K$ for DPM with Gamma kernel. Given the priors $V_j \sim \exp(\theta)$, $u_j \sim \text{Inv-Gamma}(\gamma, \beta)$ and the likelihood $X_i | c_i = j, \boldsymbol{\psi}_j \sim \text{Gamma}(V_j, V_j/u_j)$, by the Bayes' rule, we have the conditional posterior for V_j

$$\begin{aligned} p(V_j|\mathbf{X}^j, u_j) &\propto p(V_j) \prod_{k=1}^{m_j} p\left(X_k^j | V_j, u_j\right) \\ &\propto e^{-\theta V_j} \prod_{k=1}^{m_j} \frac{(V_j/u_j)^{V_j}}{\Gamma(V_j)} (X_k^j)^{V_j - 1} e^{-(V_j/u_j)X_k^j} \\ &\propto \frac{V_j^{m_j V_j}}{\Gamma(V_j)^{m_j}} \exp\left\{ -V_j \left[\theta + \frac{\sum_{k=1}^{m_j} X_k^j}{u_j} + m_j \log(u_j) - \sum_{k=1}^{m_j} \log\left(X_k^j\right) \right] \right\}. \end{aligned} \tag{17}$$

Since the conditional posterior of V_j in Equation (17) is not a standard distribution, we develop an M-H sampling algorithm to generate samples of V_j. We first find an appropriate proposal distribution for the M-H sampling. To get a fair degree of probability drawing samples from the tail part of the conditional posterior $p(V_j|\mathbf{X}^j, u_j)$, the Stirling approximation, $n! \approx \sqrt{2\pi n}(n/\mathrm{e})^n$ for large n, is used to find an appropriate family for the proposal distribution. Since $\Gamma(n) = (n-1)!$,

$$\begin{aligned} p\left(V_j|\mathbf{X}^j, u_j\right) &\propto \frac{V_j^{m_j V_j}}{\Gamma(V_j)^{m_j}} \exp\left\{ -V_j \left[\theta + \frac{\sum_{k=1}^{m_j} X_k^j}{u_j} + m_j \log(u_j) - \sum_{k=1}^{m_j} \log(X_k^j) \right] \right\} \\ &\approx \left[\frac{V_j^{V_j}}{\sqrt{2\pi(V_j-1)}\left(\frac{V_j-1}{e}\right)^{V_j-1}} \right]^{m_j} e^{-V_j B}, \text{ if } V_j \text{ is large} \\ &\approx \left[\frac{V_j^{V_j}}{(V_j-1)^{V_j}} \sqrt{\frac{V_j-1}{2\pi}} e^{V_j - 1} \right]^{m_j} e^{-V_j B} \\ &\approx \left(\sqrt{\frac{V_j-1}{2\pi}} e^{V_j} \right)^{m_j} e^{-V_j B} \approx \left(\frac{1}{2\pi} \right)^{m_j/2} (V_j)^{m_j/2} e^{-V_j(B - m_j)} \end{aligned}$$

where $B = \theta + \sum_{k=1}^{m_j} X_k^j/u_j + m_j \log(u_j) - \sum_{k=1}^{m_j} \log(X_k^j)$. This approximation holds when V_j is large and it returns a Gamma kernel function. Thus, we choose the proposal distribution to be $\text{Gamma}(r, r/V_j^0)$ with mean V_j^0 denoting the sample obtained from the previous M-H iteration. To have a non-negligible probability to draw samples far from V_j^0, the value of r is recommended to be small, e.g., $r = 2$ used in our empirical study.

Next we derive the conditional posterior for parameter u_j. By applying the Bayes' rule, we have

$$\begin{aligned} p(u_j|\mathbf{X}^j, V_j) &\propto p(u_j) \prod_{k=1}^{m_j} p(X_k^j | V_j, u_j) \\ &\propto u_j^{-(\gamma+1)} e^{-\beta/u_j} \prod_{i=1}^{m_j} \frac{(V_j/u_j)^{V_j}}{\Gamma(V_j)} (X_k^j)^{V_j - 1} e^{-(V_j/u_j)X_k^j} \end{aligned}$$

$$\propto u_j^{-(\gamma+1+m_jV_j)} \exp\left[-\frac{\beta+V_j\sum_{k=1}^{m_j}X_k^j}{u_j}\right]$$
$$\sim \text{Inv-Gamma}\left(\gamma+m_jV_j, \beta+V_j\sum_{k=1}^{m_j}X_k^j\right).$$

A.2.2. Conditional Posteriors of DPM with Gaussian Kernel For DPM with Gaussian kernel, we choose a conditional conjugate joint prior distribution for the component parameters $\boldsymbol{\psi}_j=(u_j,\sigma_j^2)$ with $j=1,2,\ldots,K$,

$$u_j|\sigma_j^2 \sim \mathcal{N}(u_0,\sigma_j^2/k_0) \text{ and } \sigma_j^2/\sigma_0^2 \sim \text{Inv-Gamma}\left(\frac{v_0}{2},\frac{1}{2}\right)$$

which determines the base function $G_0(u,\sigma^2)$ with hyper-parameters $\boldsymbol{\theta}_G=(u_0,k_0,v_0,\sigma_0)$.

Here, we derive the conditional posteriors of the latent variables $\mathbf{c}$. For $i=1,2,\ldots,m$, if X_i is associated to an existing component, by applying the Bayes' rule,

$$p(c_i=j|\mathbf{c}^{-i},\boldsymbol{\psi}_j,\alpha,X_i)=b_0\cdot p(c_i=j|\alpha,\mathbf{c}^{-i})p(X_i|c_i=j,\boldsymbol{\psi}_j)=b_0\frac{m_j^{-i}}{m+\alpha-1}\frac{1}{\sqrt{2\pi}\sigma_j}e^{-(X_i-u_j)^2/2\sigma_j^2}.$$

If X_i is associated to a new component,

$$\begin{aligned}
&p(c_i=j|\mathbf{c}^{-i},\boldsymbol{\psi}_j,\alpha,X_i)=b_0\cdot p(c_i=j|\alpha,\mathbf{c}^{-i})p(X_i|c_i=j,\boldsymbol{\psi}_j)\\
&=b_0\frac{\alpha}{m+\alpha-1}\int_0^\infty\int_{-\infty}^\infty p(X_i|u_j,\sigma_j^2)p(u_j|\sigma_j^2)p(\sigma_j^2)du_jd\sigma_j^2\\
&=b_0\frac{\alpha}{m+\alpha-1}\int_0^\infty\int_{-\infty}^\infty(2\pi\sigma_j^2)^{-1/2}e^{-\frac{(X_i-u_j)^2}{2\sigma_j^2}}\times\left(\frac{2\pi\sigma_j^2}{k_0}\right)^{-1/2}\exp\left[-\frac{k_0(u_j-u_0)^2}{2\sigma_j^2}\right]\\
&\quad\times\frac{(v_0/2)^{v_0/2}}{\Gamma(v_0/2)}\sigma_0^{v_0}(\sigma_j^2)^{-(v_0/2+1)}e^{\left(-\frac{v_0\sigma_0^2}{2\sigma_j^2}\right)}du_jd\sigma_j^2\\
&=b_0\frac{\alpha}{m+\alpha-1}\frac{(v_0/2)^{v_0/2}}{\Gamma(v_0/2)}\sigma_0^{v_0}\sqrt{k_0/2\pi}\int_0^\infty\int_{-\infty}^\infty(\sigma_j^2)^{-(\frac{v_0+3}{2})}(2\pi\sigma_j^2)^{-1/2}\\
&\quad\times\exp\left[-\left(\frac{(k_0+1)(u_j-\frac{X_i+k_0u_0}{k_0+1})^2+\frac{k_0(X_i-u_0)^2}{k_0+1}+v_0\sigma_0^2}{2\sigma_j^2}\right)\right]du_jd\sigma_j^2\\
&=b_0\frac{\alpha}{m+\alpha-1}\frac{(v_0/2)^{v_0/2}}{\Gamma(v_0/2)}\sigma_0^{v_0}\sqrt{\frac{k_0}{2\pi(k_0+1)}}\int_0^\infty\int_{-\infty}^\infty\left(\frac{2\pi\sigma_j^2}{k_0+1}\right)^{-1/2}\\
&\quad\times\exp\left[-\left(\frac{\left(u_j-\frac{X_i+k_0u_0}{k_0+1}\right)^2}{2\sigma_j^2/(k_0+1)}\right)\right]du_j\exp\left[-\left(\frac{\frac{k_0(X_i-\mu_0)^2}{k_0+1}+v_0\sigma_0^2}{2\sigma_j^2}\right)\right](\sigma_j^2)^{-(\frac{v_0+1}{2}+1)}d\sigma_j^2\\
&=b_0\frac{\alpha}{m+\alpha-1}\frac{(v_0/2)^{v_0/2}}{\Gamma(v_0/2)}\sigma_0^{v_0}\sqrt{\frac{k_0}{2\pi(k_0+1)}}\int_0^\infty\exp\left[-\left(\frac{\frac{k_0(X_i-u_0)^2}{k_0+1}+v_0\sigma_0^2}{2\sigma_j^2}\right)\right](\sigma_j^2)^{-(\frac{v_0+1}{2}+1)}d\sigma_j^2\\
&=b_0\frac{\alpha}{m+\alpha-1}\frac{(v_0/2)^{v_0/2}}{\Gamma(v_0/2)}\sigma_0^{v_0}\sqrt{\frac{k_0}{2\pi(k_0+1)}}\frac{\Gamma(A)}{B^A}
\end{aligned}$$

where b_0 is a normalization constant, $A=\frac{v_0+1}{2}$ and $B=\left[v_0\sigma_0^2+\frac{k_0(X_i-u_0)^2}{k_0+1}\right]/2$.

A.2.3. Conditional Posteriors of DPM with Beta Kernel In this section, we first find a conjugate joint prior and then derive the conditional posteriors of parameters $\boldsymbol{\psi}_j = (\gamma_j, \beta_j)$ with $j = 1, 2, \ldots, K$ for DPM with Beta kernel density. The likelihood is $X_i | c_i = j, \boldsymbol{\psi}_j \sim \text{Beta}(\gamma_j, \beta_j)$. Since Beta distribution belongs to the exponential family, we rewrite the Beta density into the general form

$$p(x|\gamma_j, \beta_j) = \frac{\Gamma(\gamma_j + \beta_j)}{\Gamma(\gamma_j)\Gamma(\beta_j)} x^{\gamma_j - 1}(1-x)^{\beta_j - 1} = \frac{\Gamma(\gamma_j + \beta_j)}{\Gamma(\gamma_j)\Gamma(\beta_j)} e^{(\gamma_j - 1)\log(x) + (\beta_j - 1)\log(1-x)}.$$

Thus, we choose a conjugate joint prior for (γ_j, β_j) with the hyper-parameters $\boldsymbol{\theta}_G = (\lambda_0, \lambda_1, \lambda_2)$ (Chick 2001)

$$\gamma_j, \beta_j | \lambda_0, \lambda_1, \lambda_2 \propto \exp\left\{ -\lambda_1 \gamma_j - \lambda_2 \beta_j - \lambda_0 \log\left[\frac{\Gamma(\gamma_j)\Gamma(\beta_j)}{\Gamma(\gamma_j + \beta_j)} \right] \right\}.$$

Then, we derive the conditional posteriors for parameters (γ_j, β_j) used in the Gibbs sampler in Appendix A.1.3. By applying the Bayes' rule, the conditional posterior for γ_j is

$$\begin{aligned}
p(\gamma_j | \beta_j, \mathbf{X}^j) &\propto p(\gamma_j | \beta_j) p(\mathbf{X}^j | \gamma_j, \beta_j) \\
&\propto \exp\left\{ -\lambda_1 \gamma_j - \lambda_0 \log\left[\frac{\Gamma(\gamma_j)}{\Gamma(\gamma_j + \beta_j)} \right] \right\} \prod_{k=1}^{m_j} \frac{\Gamma(\gamma_j + \beta_j)}{\Gamma(\gamma_j)} (X_k^j)^{\gamma_j - 1} \\
&\propto \exp\left\{ \left(-\lambda_1 + \sum_{k=1}^{m_j} \log(X_k^j) \right) \gamma_j - (\lambda_0 + m_j) \log\left[\frac{\Gamma(\gamma_j)}{\Gamma(\gamma_j + \beta_j)} \right] \right\}. \qquad (18)
\end{aligned}$$

Since the conditional posterior for γ_j in Equation (18) is not a standard distribution, we develop an M-H sampling algorithm to draw samples of γ_j by following the similar procedure used in DPM with Gamma kernel density. The Stirling approximation is used to find an appropriate proposal distribution family. As γ_j is large, the conditional posterior distribution can be approximated by

$$\begin{aligned}
p(\gamma_j | \beta_j, \mathbf{X}^j) &\propto e^{\left(-\lambda_1 + \sum_{k=1}^{m_j} \log(X_k^j) \right) \gamma_j - (\lambda_0 + m_j) \log\left[\frac{\Gamma(\gamma_j)}{\Gamma(\gamma_j + \beta_j)} \right]} \\
&\approx e^{\left(-\lambda_1 + \sum_{k=1}^{m_j} \log(X_k^j) \right) \gamma_j} \left[\frac{(\gamma_j + \beta_j - 1)!}{(\gamma_j - 1)!} \right]^{\lambda_0 + m_j} \\
&\approx e^{-\left(\lambda_1 - \sum_{k=1}^{m_j} \log(X_k^j) \right) \gamma_j} \left(\gamma_j^{\beta_j} \right)^{\lambda_0 + m_j}, \text{ if } \gamma_j \text{ is large} \\
&\sim \text{Gamma}\left(\beta_j(\lambda_0 + m_j) + 1, \lambda_1 - \sum_{k=1}^{m_j} \log(X_k^j) \right).
\end{aligned}$$

Thus, $\text{Gamma}(r, r/\gamma_j^0)$ with small r, e.g., $r = 2$ used in the empirical study, is used as the proposal distribution, where γ_j^0 denotes the sample obtained from the previous M-H sampling iteration.

Next, by applying the Bayes' rule, we derive the conditional posterior for β_j

$$\begin{aligned}
p(\beta_j | \gamma_j, \mathbf{X}^j) &\propto p(\beta_j | \gamma_j) p(\mathbf{X}^j | \gamma_j, \beta_j) \\
&\propto \exp\left\{ -\lambda_2 \beta_j - \lambda_0 \log\left[\frac{\Gamma(\beta_j)}{\Gamma(\gamma_j + \beta_j)} \right] \right\} \prod_{k=1}^{m_j} \frac{\Gamma(\gamma_j + \beta_j)}{\Gamma(\beta_j)} \left(1 - X_k^j\right)^{\beta_j - 1} \\
&\propto \exp\left\{ \left(-\lambda_2 + \sum_{k=1}^{m_j} \log(1 - X_k^j) \right) \beta_j - (\lambda_0 + m_j) \log\left[\frac{\Gamma(\beta_j)}{\Gamma(\gamma_j + \beta_j)} \right] \right\}. \qquad (19)
\end{aligned}$$

Notice that Equations (18) and (19) have the similar form, and they do not belong to any standard distribution. Thus, an M-H sampling approach is developed to generate samples for β_j. An appropriate proposal distribution family is found by applying the Stirling approximation,

$$\begin{aligned} p(\beta_j|\gamma_j, \mathbf{X}^j) &\propto e^{\left(-\lambda_2+\sum_{k=1}^{m_j}\log(1-X_k^j)\right)\beta_j-(\lambda_0+m_j)\log\left[\frac{\Gamma(\beta_j)}{\Gamma(\gamma_j+\beta_j)}\right]} \\ &\approx e^{\left(-\lambda_2+\sum_{k=1}^{m_j}\log(1-X_k^j)\right)\beta_j}\left[\frac{(\gamma_j+\beta_j-1)!}{(\beta_j-1)!}\right]^{\lambda_0+m_j} \\ &\approx e^{-\left(\lambda_2-\sum_{k=1}^{m_j}\log(1-X_k^j)\right)\beta_j}\left(\beta_j^{\gamma_j}\right)^{\lambda_0+m_j}, \text{ if } \beta_j \text{ is large} \\ &\sim \text{Gamma}\left(\gamma_j(\lambda_0+m_j)+1, \lambda_2-\sum_{k=1}^{m_j}\log(1-X_k^j)\right). \end{aligned}$$

In the M-H sampling, Gamma$(r, r/\beta_j^0)$ with small r is used as the proposal distribution, where β_j^0 denotes the sample obtained from the previous iteration.

Appendix B: Asymptotic Properties of Bayesian Nonparametric Framework

In this section we prove Theorem 2. Let $\mathcal{F}_B \equiv \{\tilde{F}^{(1)}, \tilde{F}^{(2)}, \ldots, \tilde{F}^{(B)}\}$ be the sample of distributions drawn from the posterior $p(F|\mathbf{X}_m)$. In the algorithm outlined in Section 3.5, we have ranked the sampled responses $\{\tilde{\mu}_{(b)}\}_{b=1}^B$ as $\tilde{\mu}_{(1)} < \tilde{\mu}_{(2)} < \ldots < \tilde{\mu}_{(B)}$. We can directly neglect the possibility of ties because the distributions of all $\tilde{\mu}_b$'s are continuous. Now suppose $(i_1, i_2, \ldots, i_B)$ is the permutation of integers $(1, 2, \ldots, B)$ such that $\tilde{\mu}_{i_b} = \tilde{\mu}_{(b)}$ for $b = 1, 2, \ldots, B$. In other words, i_b is the original subscript of $\tilde{\mu}_{(b)}$ before they are ranked. We define a sequence with subscript "(b)" as the same sequence with the original subscript i_b, i.e. $\mu_{(b)} = \mu_{i_b},\ \sigma_{(b)} = \sigma_{i_b},\ n_{(b)} = n_{i_b},\ S_{(b)} = S_{i_b},\ \bar{Y}_{(b)} = \bar{Y}_{i_b}$, for $b = 1, 2, \ldots, B$.

We can also rank the mean system response $\{\mu_b\}_{b=1}^B$ into $\mu_{[1]} < \mu_{[2]} < \ldots < \mu_{[B]}$, neglecting the possibility of ties. Similarly to above, suppose $(j_1, j_2, \ldots, j_B)$ is the permutation of integers $(1, 2, \ldots, B)$ such that $\mu_{j_b} = \mu_{[b]}$ for $b = 1, 2, \ldots, B$. In other words, j_b is the original subscript of $\mu_{[b]}$ before they are ranked. We define a sequence with subscript "$[b]$" as the same sequence with the original subscript j_b, i.e. $\sigma_{[b]} = \sigma_{j_b},\ n_{[b]} = n_{j_b},\ S_{[b]} = S_{j_b},\ \bar{Y}_{[b]} = \bar{Y}_{j_b},\ \tilde{\mu}_{[b]} = \tilde{\mu}_{j_b}$, for $b = 1, 2, \ldots, B$.

Finally we can rank the sample means of simulation outputs $\{\bar{Y}_b\}_{b=1}^B$ into $\bar{Y}_{\{1\}} < \bar{Y}_{\{2\}} \leq \ldots < \bar{Y}_{\{B\}}$, neglecting the possibility of ties. Similarly to above, suppose $(k_1, k_2, \ldots, k_B)$ is the permutation of integers $(1, 2, \ldots, B)$ such that $\bar{Y}_{k_b} = \bar{Y}_{\{b\}}$ for $b = 1, 2, \ldots, B$. In other words, k_b is the original subscript of $\bar{Y}_{\{b\}}$ before they are ranked. We define a sequence with subscript "$\{b\}$" as the same sequence with the original subscript k_b, i.e. $\mu_{\{b\}} = \mu_{k_b},\ \sigma_{\{b\}} = \sigma_{k_b},\ n_{\{b\}} = n_{k_b},\ S_{\{b\}} = S_{k_b},\ \tilde{\mu}_{\{b\}} = \tilde{\mu}_{k_b}$, for $b = 1, 2, \ldots, B$.

To prove the bounds in Part (i) of Theorem 2, we first show that the aforementioned three different types of ranking from $\{\mu_{[b]}\}_{b=1}^B$, $\{\bar{Y}_{\{b\}}\}_{b=1}^B$ and $\{\tilde{\mu}_{(b)}\}_{b=1}^B$ agree with each other with high probability.

LEMMA 1. *Suppose Conditions (1) and (2) of Theorem 2 hold. Then conditional on* $\mathcal{F}_B$ *and the real-world data* $\mathbf{X}_m$*, with probability at least* $1 - 1/B^2$ *for* $n_{\min} \geq (\log B)^3 / \min\left(\min_{b=1,\ldots,B-1}\left(\mu_{[b+1]} - \mu_{[b]}\right)^2, 1\right)$ *and for all large* B*, the rankings of three sequences* $\{\mu_{[b]}\}_{b=1}^B$*,* $\{\bar{Y}_{\{b\}}\}_{b=1}^B$ *and* $\{\tilde{\mu}_{(b)}\}_{b=1}^B$ *agree with each other, i.e. the permutations* $(i_1, i_2, \ldots, i_B)$*,* $(j_1, j_2, \ldots, j_B)$*, and* $(k_1, k_2, \ldots, k_B)$ *are exactly identical to each other.*

Proof of Lemma 1:

Define two events $E_1 = \{(j_1, j_2, \ldots, j_B) \text{ is identical to } (k_1, k_2, \ldots, k_B)\}$ and $E_2 = \{(i_1, i_2, \ldots, i_B) \text{ is identical to } (k_1, k_2, \ldots, k_B)\}$. We proceed in 3 steps: First, show that $\mathrm{P}(E_1) \geq 1 - 1/(2B^2)$ for all large B; Second, show that $\mathrm{P}(E_2) \geq 1 - 1/(2B^2)$ for all large B; Third, show the conclusion of the lemma.

Step 1: Show that with probability at least $1 - 1/(2B^2)$ for all large B, the ranking of $\{\bar{Y}_{\{b\}}\}_{b=1}^B$ agrees with the ranking of $\{\mu_{[b]}\}_{b=1}^B$, i.e. $\mathrm{P}(E_1) \geq 1 - 1/(2B^2)$ for all large B.

According to the definitions above, we have

$$\bar{Y}_{[b]} \;\Big|\; \mu_{[b]}, \sigma^2_{[b]} \sim \mathcal{N}\left(\mu_{[b]}, \frac{\sigma^2_{[b]}}{n_{[b]}}\right), \tag{20}$$

$$\tilde{\mu}_{[b]} \;\Big|\; \bar{Y}_{[b]}, S^2_{[b]} \sim \mathcal{N}\left(\bar{Y}_{[b]}, \frac{S^2_{[b]}}{n_{[b]}}\right). \tag{21}$$

Define new random variables $\Delta\mu_{[b]} \equiv \mu_{[b+1]} - \mu_{[b]}$ and $\Delta\bar{Y}_{[b]} \equiv \bar{Y}_{[b+1]} - \bar{Y}_{[b]}$ for all $b = 1, 2, \ldots, B$. Note that conditional on $\{\mu_{[b]}\}_{b=1}^B$ and $\{\sigma_{[b]}\}_{b=1}^B$, all $\{\bar{Y}_{[b]}\}_{b=1}^B$ are independent normal random variables. Therefore, we can obtain directly from (20) that

$$\Delta\bar{Y}_{[b]} \;\Big|\; \mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]} \sim \mathcal{N}\left(\Delta\mu_{[b]}, \frac{\sigma^2_{[b]}}{n_{[b]}} + \frac{\sigma^2_{[b+1]}}{n_{[b+1]}}\right). \tag{22}$$

Using a union bound, we can obtain that

$$\begin{aligned}
&\mathrm{P}\left(\bar{Y}_{[1]} \leq \bar{Y}_{[2]} \leq \ldots \leq \bar{Y}_{[B]} \text{ does not hold} \;\middle|\; \{\mu_{[b]}\}_{b=1}^B, \{\sigma^2_{[b]}\}_{b=1}^B\right) \\
=&\,\mathrm{P}\left(\bar{Y}_{[1]} \leq \bar{Y}_{[2]}, \bar{Y}_{[2]} \leq \bar{Y}_{[3]}, \ldots, \bar{Y}_{[B-1]} \leq \bar{Y}_{[B]} \text{ does not hold} \;\middle|\; \{\mu_{[b]}\}_{b=1}^B, \{\sigma^2_{[b]}\}_{b=1}^B\right) \\
\leq&\sum_{b=1}^{B-1} \mathrm{P}\left(\bar{Y}_{[b]} > \bar{Y}_{[b+1]} \;\middle|\; \{\mu_{[b]}\}_{b=1}^B, \{\sigma^2_{[b]}\}_{b=1}^B\right) \\
\leq&\sum_{b=1}^{B-1} \mathrm{P}\left(\left|\Delta\bar{Y}_{[b]} - \Delta\mu_{[b]}\right| \geq \Delta\mu_{[b]} \;\middle|\; \{\mu_{[b]}\}_{b=1}^B, \{\sigma^2_{[b]}\}_{b=1}^B\right) \\
\overset{(*)}{\leq}&\sum_{b=1}^{B-1} 2\exp\left(-\frac{n_{\min}(\mu_{[b+1]} - \mu_{[b]})^2}{4C_\sigma}\right) \overset{(**)}{\leq} 2B\exp\left(-\frac{(\log B)^3}{4C_\sigma}\right) \overset{(***)}{\leq} \frac{1}{2B^2}.
\end{aligned}$$

In the display above, (*) follows from the Gaussian concentration inequality (Theorem 3.4 in Massart 2007) and Condition (2), which implies that $\mathrm{Var}(\bar{Y}_{[b+1]} - \bar{Y}_{[b]}) \le 2C_\sigma/n_{\min}$; (**) is from our condition that $n_{\min} \ge (\log B)^3 / \min_{b=1,\ldots,B-1} \left(\mu_{[b+1]} - \mu_{[b]}\right)^2$; (***) follows because for B sufficiently large, $(\log B)^3/(4C_\sigma) - \log(2B) \gg 2\log B + \log 2$. This concludes the proof in Step 1.

Step 2: Show that with probability at least $1 - 1/(2B^2)$ for all large B, the rankings of $\{\tilde{\mu}_{(b)}\}_{b=1}^B$ agrees with the ranking of $\{\mu_{[b]}\}_{b=1}^B$, i.e. $\mathrm{P}(E_2) \ge 1 - 1/(2B^2)$ for all large B.

Define new random variables $\Delta\tilde{\mu}_{[b]} \equiv \tilde{\mu}_{[b+1]} - \tilde{\mu}_{[b]}$ for all $b = 1, 2, \ldots, B$. From (21) and the conditional independence between $\tilde{\mu}_{[b]}$'s, the conditional distribution of the difference $\Delta\tilde{\mu}_{[b]}$ is given by

$$\Delta\tilde{\mu}_{[b]} \;\Big|\; \bar{Y}_{[b]}, \bar{Y}_{[b+1]}, S^2_{[b]}, S^2_{[b+1]}, \mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]} \sim \mathcal{N}\left(\Delta\bar{Y}_{[b]}, \frac{S^2_{[b]}}{n_{[b]}} + \frac{S^2_{[b+1]}}{n_{[b+1]}}\right). \tag{23}$$

Define new random variables $W_{[b]} = (n_{[b+1]}/\sigma^2_{[b+1]})\bar{Y}_{[b+1]} + (n_{[b]}/\sigma^2_{[b]})\bar{Y}_{[b]}$ for $b = 1, 2, \ldots, B$. Then $(W_{[b]}, \Delta\bar{Y}_{[b]})$ is a one-to-one transformation of $(\bar{Y}_{[b+1]}, \bar{Y}_{[b]})$. Since the σ-algebra generated by $\left\{\bar{Y}_{[b]}, \bar{Y}_{[b+1]}, S^2_{[b]}, S^2_{[b+1]}, \mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\right\}$ is the same as the σ-algebra generated by $\left\{W_{[b]}, \Delta\bar{Y}_{[b]}, S^2_{[b]}, S^2_{[b+1]}, \mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\right\}$, (23) implies that

$$\Delta\tilde{\mu}_{[b]} \;\Big|\; W_{[b]}, \Delta\bar{Y}_{[b]}, S^2_{[b]}, S^2_{[b+1]}, \mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]} \sim \mathcal{N}\left(\Delta\bar{Y}_{[b]}, \frac{S^2_{[b]}}{n_{[b]}} + \frac{S^2_{[b+1]}}{n_{[b+1]}}\right). \tag{24}$$

From (20), we know that conditional on $\mu_{[b]}$ and $\sigma_{[b]}$, the sample mean $\bar{Y}_{[b]}$ is a normal random variable. Furthermore, conditional on $\{\mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\}$, we have that $\bar{Y}_{[b]}$ and $\bar{Y}_{[b+1]}$ are also independent normal random variables. Since $(W_{[b]}, \Delta\bar{Y}_{[b]})$ is a one-to-one transformation of $(\bar{Y}_{[b+1]}, \bar{Y}_{[b]})$, we have that conditional on $\{\mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\}$, $W_{[b]}$ and $\Delta\bar{Y}_{[b]}$ are also jointly normal, with the conditional covariance

$$\begin{aligned}
&\mathrm{Cov}\left(W_{[b]}, \Delta\bar{Y}_{[b]} \;\middle|\; \mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\right) \\
&= \mathrm{Cov}\left((n_{[b+1]}/\sigma^2_{[b+1]})\bar{Y}_{[b+1]} + (n_{[b]}/\sigma^2_{[b]})\bar{Y}_{[b]},\ \bar{Y}_{[b+1]} - \bar{Y}_{[b]} \;\middle|\; \mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\right) \\
&= \frac{n_{[b+1]}}{\sigma^2_{[b+1]}}\mathrm{Var}\left(\bar{Y}_{[b+1]} \;\middle|\; \mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\right) - \frac{n_{[b]}}{\sigma^2_{[b]}}\mathrm{Var}\left(\bar{Y}_{[b]} \;\middle|\; \mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\right) \\
&= \frac{n_{[b+1]}}{\sigma^2_{[b+1]}} \cdot \frac{\sigma^2_{[b+1]}}{n_{[b+1]}} - \frac{n_{[b]}}{\sigma^2_{[b]}} \cdot \frac{\sigma^2_{[b]}}{n_{[b]}} = 0.
\end{aligned}$$

Hence $W_{[b]}$ and $\Delta\bar{Y}_{[b]}$ are also independent normal random variables conditional on $\{\mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\}$. Since $\bar{Y}_{[b]}$ and $\bar{Y}_{[b+1]}$ are sample means, conditional on $\{\mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\}$, we have that the random vector $(\bar{Y}_{[b+1]}, \bar{Y}_{[b]})$ is independent of the sample variances $(S^2_{[b]}, S^2_{[b+1]})$.

Hence $(W_{[b]}, \Delta\bar{Y}_{[b]})$ is also independent of the sample variances $(S^2_{[b]}, S^2_{[b+1]})$ conditional on $\{\mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\}$. In other words, for the joint density of $W_{[b]}$ and $\Delta\bar{Y}_{[b]}$, we have

$$\begin{aligned}
&p\left(w_{[b]}, \Delta\bar{y}_{[b]} \;\middle|\; S^2_{[b]}, S^2_{[b+1]}, \mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\right) \\
=&\, p\left(w_{[b]}, \Delta\bar{y}_{[b]} \;\middle|\; \mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\right) \\
=&\, p\left(\Delta\bar{y}_{[b]} \;\middle|\; \mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\right) \cdot p\left(w_{[b]} \;\middle|\; \mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\right) \qquad (25)
\end{aligned}$$

For short let $v_{1b} = \frac{\sigma^2_{[b]}}{n_{[b]}} + \frac{\sigma^2_{[b+1]}}{n_{[b+1]}}$ and $v_{2b} = \frac{S^2_{[b]}}{n_{[b]}} + \frac{S^2_{[b+1]}}{n_{[b+1]}}$. By combining (22), (24) and (25) together, we can marginalize out $W_{[b]}$ and $\Delta\bar{Y}_{[b]}$, and obtain that

$$\begin{aligned}
&p\left(\Delta\tilde{\mu}_{[b]} \;\middle|\; S^2_{[b]}, S^2_{[b+1]}, \mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\right) \\
=&\iint p\left(\Delta\tilde{\mu}_{[b]} \;\middle|\; w_{[b]}, \Delta\bar{y}_{[b]}, S^2_{[b]}, S^2_{[b+1]}, \mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\right) \\
&\cdot p\left(w_{[b]}, \Delta\bar{y}_{[b]} \;\middle|\; S^2_{[b]}, S^2_{[b+1]}, \mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\right) d\Delta\bar{y}_{[b]} dw_{[b]} \\
=&\iint p\left(\Delta\tilde{\mu}_{[b]} \;\middle|\; w_{[b]}, \Delta\bar{y}_{[b]}, S^2_{[b]}, S^2_{[b+1]}, \mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\right) \\
&\cdot p\left(\Delta\bar{y}_{[b]} \;\middle|\; \mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\right) \cdot p\left(w_{[b]} \;\middle|\; \mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\right) d\Delta\bar{y}_{[b]} dw_{[b]} \\
\overset{(*)}{=}&\int p\left(\Delta\tilde{\mu}_{[b]} \;\middle|\; w_{[b]}, \Delta\bar{y}_{[b]}, S^2_{[b]}, S^2_{[b+1]}, \mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\right) \cdot p\left(\Delta\bar{y}_{[b]} \;\middle|\; \mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}\right) d\Delta\bar{y}_{[b]} \\
\propto&\int \exp\left\{-\frac{\left(\Delta\tilde{\mu}_{[b]} - \Delta\bar{y}_{[b]}\right)^2}{2v_{2b}}\right\} \cdot \exp\left\{-\frac{\left(\Delta\bar{y}_{[b]} - \Delta\mu_{[b]}\right)^2}{2v_{1b}}\right\} d\Delta\bar{y}_{[b]} \\
=&\int \exp\left\{-\frac{\left(v_{1b}^{-1} + v_{2b}^{-1}\right)\left(\Delta\bar{y}_{[b]}\right)^2}{2} + \left(v_{1b}^{-1}\Delta\mu_{[b]} + v_{2b}^{-1}\Delta\tilde{\mu}_{[b]}\right)\Delta\bar{y}_{[b]} - \frac{\left(\Delta\mu_{[b]}\right)^2}{2v_{1b}} - \frac{\left(\Delta\tilde{\mu}_{[b]}\right)^2}{2v_{2b}}\right\} d\Delta\bar{y}_{[b]} \\
=&\int \exp\left\{-\frac{\left(v_{1b}^{-1} + v_{2b}^{-1}\right)}{2}\left[\Delta\bar{y}_{[b]} - \frac{v_{1b}^{-1}\Delta\mu_{[b]} + v_{2b}^{-1}\Delta\tilde{\mu}_{[b]}}{v_{1b}^{-1} + v_{2b}^{-1}}\right]^2\right\} d\Delta\bar{y}_{[b]} \\
&\cdot \exp\left\{\frac{\left(v_{1b}^{-1}\Delta\mu_{[b]} + v_{2b}^{-1}\Delta\tilde{\mu}_{[b]}\right)^2}{2\left(v_{1b}^{-1} + v_{2b}^{-1}\right)} - \frac{\left(\Delta\mu_{[b]}\right)^2}{2v_{1b}} - \frac{\left(\Delta\tilde{\mu}_{[b]}\right)^2}{2v_{2b}}\right\} \\
\overset{(**)}{=}&\exp\left\{\frac{\left(v_{1b}^{-1}\Delta\mu_{[b]} + v_{2b}^{-1}\Delta\tilde{\mu}_{[b]}\right)^2}{2\left(v_{1b}^{-1} + v_{2b}^{-1}\right)} - \frac{\left(\Delta\mu_{[b]}\right)^2}{2v_{1b}} - \frac{\left(\Delta\tilde{\mu}_{[b]}\right)^2}{2v_{2b}}\right\} \\
=&\exp\left\{-\frac{\left(\Delta\tilde{\mu}_{[b]}\right)^2}{2(v_{1b} + v_{2b})} + \frac{\Delta\mu_{[b]} \cdot \Delta\tilde{\mu}_{[b]}}{v_{1b} + v_{2b}} - \frac{\left(\Delta\mu_{[b]}\right)^2}{2(v_{1b} + v_{2b})}\right\} \\
=&\exp\left\{-\frac{\left(\Delta\tilde{\mu}_{[b]} - \Delta\mu_{[b]}\right)^2}{2(v_{1b} + v_{2b})}\right\},
\end{aligned}$$

where (*) follows because the integral with respect to $w_{[b]}$ is equal to 1, and (**) follows because the integral with respect to $\bar{y}_{[b]}$ is a Gaussian integral and is equal to a constant. Therefore, we have shown that

$$\Delta\tilde{\mu}_{[b]} \;\middle|\; \mu_{[b]}, \mu_{[b+1]}, \sigma^2_{[b]}, \sigma^2_{[b+1]}, S^2_{[b]}, S^2_{[b+1]} \sim \mathcal{N}\left(\Delta\mu_{[b]}, \frac{\sigma^2_{[b]}}{n_{[b]}} + \frac{\sigma^2_{[b+1]}}{n_{[b+1]}} + \frac{S^2_{[b]}}{n_{[b]}} + \frac{S^2_{[b+1]}}{n_{[b+1]}}\right).$$

Similar to the proof in Step 1, we can control the probability of incorrect ranking:

$$\begin{aligned}
&\mathrm{P}\left(\tilde{\mu}_{[1]} \le \tilde{\mu}_{[2]} \le \ldots \le \tilde{\mu}_{[B]} \text{ does not hold} \;\middle|\; \{S^2_{[b]}\}_{b=1}^B, \{\mu_{[b]}\}_{b=1}^B, \{\sigma^2_{[b]}\}_{b=1}^B\right) \\
\le &\mathrm{P}\left(\tilde{\mu}_{[1]} \le \tilde{\mu}_{[2]}, \ldots, \tilde{\mu}_{[B-1]} \le \tilde{\mu}_{[B]} \text{ does not hold} \;\middle|\; \{S^2_{[b]}\}_{b=1}^B, \{\mu_{[b]}\}_{b=1}^B, \{\sigma^2_{[b]}\}_{b=1}^B\right) \\
\le &\sum_{b=1}^{B-1} \mathrm{P}\left(\tilde{\mu}_{[b]} > \tilde{\mu}_{[b+1]} \;\middle|\; \{S^2_{[b]}\}_{b=1}^B, \{\mu_{[b]}\}_{b=1}^B, \{\sigma^2_{[b]}\}_{b=1}^B\right) \\
\le &\sum_{b=1}^{B-1} \mathrm{P}\left(\left|\Delta\tilde{\mu}_{[b]} - \Delta\mu_{[b]}\right| > \Delta\mu_{[b]} \;\middle|\; \{S^2_{[b]}\}_{b=1}^B, \{\mu_{[b]}\}_{b=1}^B, \{\sigma^2_{[b]}\}_{b=1}^B\right) \\
\overset{(*)}{\le} &2\sum_{b=1}^{B-1} \exp\left(-\frac{n_{\min}(\mu_{[b+1]} - \mu_{[b]})^2}{2\left(2C_\sigma + S^2_{[b]} + S^2_{[b+1]}\right)}\right),
\end{aligned} \tag{26}$$

where (*) comes from the Gaussian concentration inequality (Theorem 3.4 in Massart 2007) and the fact that $n_{\min} \le n_b$ for all $b = 1, 2, \ldots, B$.

Next we remove the conditioning on $\{S^2_{[b]}\}_{b=1}^B$ in (26).

$$\begin{aligned}
&\mathrm{P}\left(\tilde{\mu}_{[1]} \le \tilde{\mu}_{[2]} \le \ldots \le \tilde{\mu}_{[B]} \text{ does not hold} \;\middle|\; \{\mu_{[b]}\}_{b=1}^B, \{\sigma^2_{[b]}\}_{b=1}^B\right) \\
= &\mathrm{E}_{\{S^2_{[b]}\}_{b=1}^B \left| \{\mu_{[b]}\}_{b=1}^B, \{\sigma^2_{[b]}\}_{b=1}^B \right.} \mathrm{P}\left(\tilde{\mu}_{[1]} \le \tilde{\mu}_{[2]} \le \ldots \le \tilde{\mu}_{[B]} \text{ does not hold} \;\middle|\; \{S^2_{[b]}\}_{b=1}^B, \{\mu_{[b]}\}_{b=1}^B, \{\sigma^2_{[b]}\}_{b=1}^B\right) \\
\le &\mathrm{E}_{\{S^2_{[b]}\}_{b=1}^B \left| \{\mu_{[b]}\}_{b=1}^B, \{\sigma^2_{[b]}\}_{b=1}^B \right.} 2\sum_{b=1}^{B-1} \exp\left(-\frac{n_{\min}(\mu_{[b+1]} - \mu_{[b]})^2}{2\left(2C_\sigma + S^2_{[b]} + S^2_{[b+1]}\right)}\right) \\
\le &\mathrm{E}_{\{S^2_{[b]}\}_{b=1}^B \left| \{\mu_{[b]}\}_{b=1}^B, \{\sigma^2_{[b]}\}_{b=1}^B \right.} 2\sum_{b=1}^{B-1} \exp\left(-\frac{n_{\min}(\mu_{[b+1]} - \mu_{[b]})^2}{2\left(2C_\sigma + S^2_{[b]} + S^2_{[b+1]}\right)}\right) \times \\
&\left[I(S^2_{[b]} \le C_1\sigma^2_{[b]}, S^2_{[b+1]} \le C_1\sigma^2_{[b+1]}) + I(S^2_{[b]} > C_1\sigma^2_{[b]}) + I(S^2_{[b+1]} > C_1\sigma^2_{[b+1]})\right] \\
\le &2B\exp\left(-\frac{n_{\min}\min_{b=1,\ldots,B}(\mu_{[b+1]} - \mu_{[b]})^2}{4C_\sigma(1+C_1)}\right) + 4\sum_{b=1}^{B} \mathrm{P}\left(S^2_{[b]} > C_1\sigma^2_{[b]} \middle| \{\mu_{[b]}\}_{b=1}^B, \{\sigma^2_{[b]}\}_{b=1}^B\right),
\end{aligned} \tag{27}$$

where $I(\cdot)$ is the indicator function and the constant C_1 is chosen to satisfy $C_1 > 5$. The first term in (27) is summable for $n_{\min} = 1, 2, \ldots$. To control the second term in (27), since $(n_{[b]} - 1)S^2_{[b]}/\sigma^2_{[b]}$ follows the chi-square distribution with $n_{[b]} - 1$ degrees of freedom, we have

$$\begin{aligned}
&\mathrm{P}\left(S^2_{[b]} > C_1\sigma^2_{[b]} \;\middle|\; \{\mu_{[b]}\}_{b=1}^B, \{\sigma^2_{[b]}\}_{b=1}^B\right) = \mathrm{P}\left(\frac{S^2_{[b]}}{\sigma^2_{[b]}} - 1 > C_1 - 1 \;\middle|\; \{\mu_{[b]}\}_{b=1}^B, \{\sigma^2_{[b]}\}_{b=1}^B\right) \\
= &\mathrm{P}\left(\frac{1}{n_{[b]}-1}\left(\frac{(n_{[b]}-1)S^2_{[b]}}{\sigma^2_{[b]}}\right) - 1 > C_1 - 1 \;\middle|\; \{\mu_{[b]}\}_{b=1}^B, \{\sigma^2_{[b]}\}_{b=1}^B\right) \\
\overset{(*)}{\le} &\exp\left(-\frac{(C_1-1)n_{[b]}}{4}\right) \le \exp\left(-\frac{(C_1-1)n_{\min}}{4}\right),
\end{aligned} \tag{28}$$

where (*) follows from $C_1 > 5$ and the concentration inequality for chi-square random variables (Lemma 1 in Laurent and Massart 2000). Based on the condition on $n_{\min}$, we can combine (27) and (28) and obtain

$$\begin{aligned}
&\mathrm{P}\left(\tilde{\mu}_{[1]} \le \tilde{\mu}_{[2]} \le \ldots \le \tilde{\mu}_{[B]} \text{ does not hold } \middle|\ \{\mu_{[b]}\}_{b=1}^{B}, \{\sigma^2_{[b]}\}_{b=1}^{B}\right) \\
&\le 2B\exp\left(-\frac{(\log B)^3}{4C_\sigma(1+C_1)}\right) + 4B\exp\left(-\frac{(C_1-1)(\log B)^3}{4}\right) \\
&\overset{(**)}{\le} \frac{1}{4B^2} + \frac{1}{4B^2} = \frac{1}{2B^2},
\end{aligned}$$

where (**) follows because $(\log B)^3/[4C_\sigma(1+C_1)] - \log(2B) \gg 2\log B + \log 4$ and $(C_1 - 1)(\log B)^3/4 - \log(4B) \gg 2\log B + \log 4$. This concludes the proof in Step 2.

Step 3: The conclusion of Lemma 1 follows trivially from Step 1 and Step 2: If both E_1 and E_2 happen, then the rankings of $\{\mu_{[b]}\}_{b=1}^{B}$, $\{\bar{Y}_{\{b\}}\}_{b=1}^{B}$ and $\{\tilde{\mu}_{(b)}\}_{b=1}^{B}$ agree with each other. Therefore, the probability that their rankings disagree is upper bounded by $\mathrm{P}(E_1^c \cup E_2^c) \le \mathrm{P}(E_1^c) + \mathrm{P}(E_2^c) \le 1/(2B^2) + 1/(2B^2) = 1/B^2$. Thus the conclusion of the lemma is proved. □

Proof of Theorem 2 (i):

Similar to the derivation of the conditional distribution of $\Delta\tilde{\mu}_{[b]}$ in the proof of Lemma 1, on Θ_B for every fixed b with $b = 1, 2, \ldots, B$, we can combine (20) and (21) and marginalize out $\bar{Y}_{[b]}$ to obtain that

$$\tilde{\mu}_{[b]} \ \middle|\ S^2_{[b]}, \mu_{[b]}, \sigma^2_{[b]} \sim \mathcal{N}\left(\mu_{[b]}, \frac{\sigma^2_{[b]} + S^2_{[b]}}{n_{[b]}}\right). \tag{29}$$

Note that the σ-algebra generated by $\{\mu_{[b]}, \sigma^2_{[b]}\}$ is included in the σ-algebra generated by $\mathcal{F}_B$. On the other hand, in (20), our simulation mechanism indicates that $\bar{Y}_{[b]}$ is a normal random variable that depends on only the value of $\{\mu_{[b]}, \sigma^2_{[b]}\}$, but not any other information in the σ-algebra generated by $\mathcal{F}_m$ and $\mathbf{X}_m$. Therefore, the conditional distribution of $\bar{Y}_{[b]}$ given $\{\mathcal{F}_B, \mathbf{X}_m\}$ is still $\mathcal{N}\left(\mu_{[b]}, \sigma^2_{[b]}/n_{[b]}\right)$, which further implies that (29) can be rewritten as

$$\tilde{\mu}_{[b]} \ \middle|\ S^2_{[b]}, \mathcal{F}_B, \mathbf{X}_m \sim \mathcal{N}\left(\mu_{[b]}, \frac{\sigma^2_{[b]} + S^2_{[b]}}{n_{[b]}}\right).$$

For sufficiently large C_1, $B \ge 2$ and any fixed $b \in \{1, 2, \ldots, B\}$,

$$\begin{aligned}
&\mathrm{P}\left(\sqrt{n_{[b]}}\left|\tilde{\mu}_{[b]} - \mu_{[b]}\right| > C_1 \log B \,\middle|\, \mathcal{F}_B, \mathbf{X}_m\right) \\
&= \mathrm{E}_{S^2_{[b]}|\mathcal{F}_B,\mathbf{X}_m} \mathrm{P}\left(\sqrt{n_{[b]}}\left|\tilde{\mu}_{[b]} - \mu_{[b]}\right| > C_1 \log B \ \middle|\ S^2_{[b]}, \mathcal{F}_B, \mathbf{X}_m\right)
\end{aligned}$$

$$
\begin{aligned}
&\overset{(*)}{\leq} \mathrm{E}_{S^2_{[b]}|\sigma^2_{[b]}} 2\exp\left(-\frac{(C_1\log B)^2}{2\left(\sigma^2_{[b]}+S^2_{[b]}\right)}\right)\\
&= \mathrm{E}_{S^2_{[b]}|\sigma^2_{[b]}} 2\exp\left(-\frac{(C_1\log B)^2}{2\left(\sigma^2_{[b]}+S^2_{[b]}\right)}\right)\cdot\left[I\left(S^2_{[b]}\leq (C_1\log B)\sigma^2_{[b]}\right)+I\left(S^2_{[b]}> (C_1\log B)\sigma^2_{[b]}\right)\right]\\
&\leq 2\exp\left(-\frac{(C_1\log B)^2}{2C_\sigma(1+C_1\log B)}\right)+2\mathrm{P}\left(S^2_{[b]}> (C_1\log B)\sigma^2_{[b]}\;\middle|\;\mu_{[b]},\sigma^2_{[b]}\right)\\
&\overset{(**)}{\leq} 2\exp\left(-\frac{C_1\log B}{4C_\sigma}\right)+2\exp\left(-\frac{(C_1\log B-1)n_{\min}}{4}\right)\overset{(***)}{\leq}\exp\left(-\frac{C_1\log B}{8C_\sigma}\right). \qquad (30)
\end{aligned}
$$

In the display above, (*) follows from the Gaussian concentration inequality (Theorem 3.4 in Massart 2007), and the conditional expectation $\mathrm{E}_{S^2_{[b]}|\mathcal{F}_B,\mathbf{X}_m}$ is equivalent to the conditional expectation $\mathrm{E}_{S^2_{[b]}|\sigma^2_{[b]}}$ because the conditional distribution of the sample variance $S^2_{[b]}$ given $\{\mathcal{F}_B,\mathbf{X}_m\}$ only depends on $\sigma^2_{[b]}$, as $(n_{[b]}-1)S^2_{[b]}/\sigma^2_{[b]}$ follows the chi-square distribution with $n_{[b]}-1$ degrees of freedom; (**) follows from (28) and sufficiently large C_1; (***) follows because by our condition on $n_{\min}$, we have $n_{\min}\to\infty$ as $B\to\infty$, $C_1\log B/(4C_\sigma)\ll(C_1\log B-1)n_{\min}/4$ and $C_1\log B/(8C_\sigma)\gg\log 4$.

Using the upper bound in (30) for a single $b\in\{1,2,\ldots,B\}$, we can derive that for sufficiently large $C_1>0$,

$$
\begin{aligned}
&\mathrm{P}\left(\max_{b=1,\ldots,B}\sqrt{n_{[b]}}\left|\tilde{\mu}_{[b]}-\mu_{[b]}\right|> C_1\log B\;\middle|\;\mathcal{F}_B,\mathbf{X}_m\right)\\
&=1-\prod_{b=1}^{B}\left[1-\mathrm{P}\left(\sqrt{n_{[b]}}\left|\tilde{\mu}_{[b]}-\mu_{[b]}\right|> C_1\log B\;\middle|\;\mathcal{F}_B,\mathbf{X}_m\right)\right]\\
&\leq 1-\left[1-\exp\left(-\frac{C_1\log B}{8C_\sigma}\right)\right]^B\overset{(*)}{\leq}-B\log\left[1-\exp\left(-\frac{C_1\log B}{8C_\sigma}\right)\right]\\
&\overset{(**)}{\leq} 2B\exp\left(-\frac{C_1\log B}{8C_\sigma}\right)=\exp\left(-\frac{C_1\log B}{8C_\sigma}+\log(2B)\right)\overset{(***)}{\leq}\frac{2}{B^2},
\end{aligned}
$$

where (*) uses $1-e^{-t}\leq t$ for $t>0$; (**) is because $-\log(1-t)\leq 2t$ for $t\in(0,1/3)$; (***) follows if we choose C_1 sufficiently large such that $C_1\log B/(8C_\sigma)>\log 3$. This implies that

$$
\mathrm{P}\left(\max_{b=1,\ldots,B}\sqrt{n_{[b]}}\left|\tilde{\mu}_{[b]}-\mu_{[b]}\right|\leq C_1\log B\;\middle|\;\mathcal{F}_B,\mathbf{X}_m\right)\geq 1-\frac{2}{B^2}, \qquad (31)
$$

for the constant C_1 chosen above, for all large B.

Now we are going to replace the difference $\sqrt{n_{[b]}}\left|\tilde{\mu}_{[b]}-\mu_{[b]}\right|$ in (31) by $\sqrt{n_{(b)}}\left|\tilde{\mu}_{(b)}-\mu_{(b)}\right|$. To do this, we need the ranking of $\{\tilde{\mu}_{(b)}\}_{b=1}^B$ to agree with the ranking of $\{\mu_{[b]}\}_{b=1}^B$. From Lemma 1, we already know that this event happens with large probability. Specifically, let Θ_B be the event that the rankings of $\{\tilde{\mu}_{(b)}\}_{b=1}^B,\{\bar{Y}_{\{b\}}\}_{b=1}^B,\{\mu_{[b]}\}_{b=1}^B$ agree with each other. Then Lemma 1 says that $\mathrm{P}(\Theta_B\mid\mathcal{F}_B,\mathbf{X}_m)\geq 1-1/B^2$ for $n_{\min}\geq(\log B)^3/\min\left(\min_{b=1,\ldots,B-1}\left(\mu_{[b+1]}-\mu_{[b]}\right)^2,1\right)$ and for all large B. It is clear that these events have the relation

$$
\left\{\max_{b=1,\ldots,B}\sqrt{n_{(b)}}\left|\tilde{\mu}_{(b)}-\mu_{(b)}\right|\leq C_1\log B\right\}\supseteq\Theta_B\cap\left\{\max_{b=1,\ldots,B}\sqrt{n_{[b]}}\left|\tilde{\mu}_{[b]}-\mu_{[b]}\right|\leq C_1\log B\right\}.
$$

Therefore, using Lemma 1, Equation (31) and a union bound, we have

$$\begin{aligned}
&\mathrm{P}\left(\max_{b=1,\ldots,B}\sqrt{n_{(b)}}\left|\tilde{\mu}_{(b)}-\mu_{(b)}\right|>C_1\log B\;\middle|\;\mathcal{F}_B,\mathbf{X}_m\right)\\
\le&\,\mathrm{P}(\Theta_B^c)+\mathrm{P}\left(\max_{b=1,\ldots,B}\sqrt{n_{[b]}}\left|\tilde{\mu}_{[b]}-\mu_{[b]}\right|\ge C_1\log B\;\middle|\;\mathcal{F}_B,\mathbf{X}_m\right)\\
\le&\,\frac{1}{B^2}+\frac{2}{B^2}=\frac{3}{B^2},
\end{aligned}$$

which implies that

$$\mathrm{P}\left(\max_{b=1,\ldots,B}\sqrt{n_{(b)}}\left|\tilde{\mu}_{(b)}-\mu_{(b)}\right|\le C_1\log B\;\middle|\;\mathcal{F}_B,\mathbf{X}_m\right)\ge 1-\frac{3}{B^2}. \tag{32}$$

In (32), we can set $b=\lceil\gamma B\rceil$ (which is an index changing with B) for a fixed $\gamma\in(0,1)$ and obtain that for $n_{\min}\ge(\log B)^3/\min\left(\min_{b=1,\ldots,B-1}\left(\mu_{[b+1]}-\mu_{[b]}\right)^2,1\right)$ and for all large B,

$$\left|\tilde{\mu}_{(\lceil\gamma B\rceil)}-\mu_{(\lceil\gamma B\rceil)}\right|=O_p\left(\frac{\log B}{\sqrt{n_{\min}}}\right), \tag{33}$$

where O_p is the conditional measure of the simulation outputs given $\mathcal{F}_B$ and the input data $\mathbf{X}_m$. Since $n_{\min}\ge(\log B)^3$ by the condition on $n_{\min}$, we have $\log B/\sqrt{n_{\min}}\le 1/\sqrt{\log B}\to 0$ as $B\to\infty$.

Next we turn to the relation between μ_b and the quantiles of $\mathrm{P}_\mu(\cdot|\mathbf{X}_m)$. For a fixed $\gamma\in(0,1)$, by the standard central limit theorem for quantiles, the sample quantile $\mu_{(\lceil\gamma B\rceil)}$ satisfies

$$\sqrt{B}\left(\mu_{(\lceil\gamma B\rceil)}-q_\gamma(\mathbf{X}_m,\mu(\cdot))\right)\xrightarrow{d}\mathcal{N}\left(0,\frac{\gamma(1-\gamma)}{p_\mu^2(q_\gamma(\mathbf{X}_m,\mu(\cdot))|\mathbf{X}_m)}\right),$$

as $B\to\infty$, where $\xrightarrow{d}$ represents the convergence in distribution, and $p_\mu(\cdot|\mathbf{X}_m)$ denotes the posterior density of $\mu(F)$ given $\mathbf{X}_m$. Condition (1) in Theorem 2 guarantees that for any $\gamma\in(0,1)$, $p_\mu(q_\gamma(\mathbf{X}_m,\mu(\cdot))|\mathbf{X}_m)>0$ so the variance in the limiting normal distribution exists. Therefore, conditional on $\mathbf{X}_m$, $\left|\mu_{(\lceil\gamma B\rceil)}-q_\gamma(\mathbf{X}_m,\mu(\cdot))\right|=O_p\left(1/\sqrt{B}\right)$. If we rewrite this relation conditional on both $\mathcal{F}_B$ and $\mathbf{X}_m$, then as $B\to\infty$, with high probability in $\mathcal{F}_B$ and $\mathbf{X}_m$,

$$\left|\mu_{(\lceil\gamma B\rceil)}-q_\gamma(\mathbf{X}_m,\mu(\cdot))\right|=O_p\left(\frac{1}{\sqrt{B}}\right). \tag{34}$$

We combine (33) and (34) by the triangular inequality to obtain that

$$\left|\tilde{\mu}_{(\lceil\gamma B\rceil)}-q_\gamma(\mathbf{X}_m,\mu(\cdot))\right|=O_p\left(\frac{\log B}{\sqrt{n_{\min}}}\right)+O_p\left(\frac{1}{\sqrt{B}}\right) \tag{35}$$

for $n_{\min}\ge(\log B)^3/\min\left(\min_{b=1,\ldots,B-1}\left(\mu_{[b+1]}-\mu_{[b]}\right)^2,1\right)$ and for all large B.

Since the Hausdorff distance between two intervals has the simple expression $d_H([a_1,b_1],[a_2,b_2])=\max(|a_1-a_2|,|b_1-b_2|)$, we can set $\gamma=\alpha^*/2$ and $\gamma=1-\alpha^*/2$ respectively in (35) and obtain that conditional on $\mathcal{F}_B$ and $\mathbf{X}_m$,

$$\begin{aligned}
&d_H\left(\left[\tilde{\mu}_{(\lceil(\alpha^*/2)B\rceil)},\tilde{\mu}_{(\lceil(1-\alpha^*/2)B\rceil)}\right],\left[q_{\alpha^*/2}(\mathbf{X}_m,\mu(\cdot)),q_{1-\alpha^*/2}(\mathbf{X}_m,\mu(\cdot))\right]\right)\\
=&\,\max\left\{\left|\tilde{\mu}_{(\lceil\alpha^*B\rceil)}-q_{\alpha^*/2}(\mathbf{X}_m,\mu(\cdot))\right|,\left|\tilde{\mu}_{(\lceil(1-\alpha^*/2)B\rceil)}-q_{1-\alpha^*/2}(\mathbf{X}_m,\mu(\cdot))\right|\right\}\\
=&\,O_p\left(\frac{\log B}{\sqrt{n_{\min}}}\right)+O_p\left(\frac{1}{\sqrt{B}}\right).
\end{aligned}$$

□

Proof of Theorem 2 (ii): Since $n_{\min} \geq (\log B)^3$, we have $\log B/\sqrt{n_{\min}} \leq 1/\sqrt{\log B} \to 0$ as $B \to \infty$. The relation (33) implies that for $n_{\min} \geq (\log B)^3 / \min\left(\min_{b=1,\ldots,B-1}\left(\mu_{[b+1]} - \mu_{[b]}\right)^2, 1\right)$ and for all large B,

$$\left|\tilde{\mu}_{(\lceil \gamma B \rceil)} - \mu_{(\lceil \gamma B \rceil)}\right| \to 0 \tag{36}$$

in probability as $B \to \infty$. Besides, from the derivation of (34), we also have the convergence of sample quantiles,

$$\left|\mu_{(\lceil \gamma B \rceil)} - q_\gamma(\mathbf{X}_m, \mu(\cdot))\right| \to 0 \tag{37}$$

in probability as $B \to \infty$.

Conditions (3) and (4) from Theorem 2 lead to the posterior consistency of $\mu(F)$, i.e. the posterior distribution $\mathrm{P}_\mu(\cdot|\mathbf{X}_m)$ converges to the point mass at the true system mean response $\mu(F^c)$ as $m \to \infty$. As a result, we have

$$|q_\gamma(\mathbf{X}_m, \mu(\cdot)) - \mu(F^c)| \to 0 \tag{38}$$

in probability as $m \to \infty$.

We combine (36)-(38), and then by the triangular inequality, we obtain

$$\left|\tilde{\mu}_{(\lceil \gamma B \rceil)} - \mu(F^c)\right| \to 0$$

in probability as $B, m \to \infty$. By setting $\gamma = \alpha^*/2, 1-\alpha^*/2$, we conclude that the CrI in (9) converges to $\mu(F^c)$ for $n_{\min} \geq (\log B)^3 / \min\left(\min_{b=1,\ldots,B-1}\left(\mu_{[b+1]} - \mu_{[b]}\right)^2, 1\right)$ as in probability $m, B \to \infty$. □

Appendix C: Variance Decomposition of System Performance Estimation

Proof of Theorem 3: Given the real-world data $\mathbf{X}_m$ and the simulation outputs $\mathcal{Y}_\mathbf{n}$, the variance of compound random variable $U = \tilde{\mu}(\tilde{F}^{(b)})$ with $\tilde{F}^{(b)} \sim p(F|\mathbf{X}_m)$ quantifies the overall uncertainty of our belief on the system performance $\mu^c = \mu(F^c)$. Here, we decompose this variance to measure the relative contributions from the input and simulation uncertainty. For notational simplification, we drop the conditional on $\mathbf{X}_m$.

At any $\tilde{F}^{(b)} \sim p(F|\mathbf{X}_m)$, let $\mu_b = \mu(\tilde{F}^{(b)})$, $\sigma_b^2 = \sigma_\epsilon^2(\tilde{F}^{(b)})$, $\tilde{\mu}_b = \tilde{\mu}(\tilde{F}^{(b)})$, $\bar{Y}_b = \sum_{j=1}^{n_b} Y_j(\tilde{F}^{(b)})/n_b$ and $\mathbf{Y}_b = \{Y_1(\tilde{F}^{(b)}), Y_2(\tilde{F}^{(b)}), \ldots, Y_{n_b}(\tilde{F}^{(b)})\}$. The variance of U can be written as

$$\begin{aligned}
&\mathrm{Var}\left[\tilde{\mu}\left(\tilde{F}^{(b)}\right)\right] = \mathrm{E}\left[\left(\tilde{\mu}_b - \mathrm{E}\left[\tilde{\mu}_b\right]\right)^2\right] \\
&= \mathrm{E}\left\{\left[\left(\tilde{\mu}_b - \bar{Y}_b\right) + \left(\bar{Y}_b - \mu_b\right) + \left(\mu_b - \mathrm{E}\left[\tilde{\mu}_b\right]\right)\right]^2\right\} \\
&= \mathrm{E}\left[\left(\tilde{\mu}_b - \bar{Y}_b\right)^2\right] + \mathrm{E}\left[\left(\bar{Y}_b - \mu_b\right)^2\right] + \mathrm{E}\left[\left(\mu_b - \mathrm{E}\left[\tilde{\mu}_b\right]\right)^2\right] + 2\mathrm{E}\left[\left(\tilde{\mu}_b - \bar{Y}_b\right)\left(\bar{Y}_b - \mu_b\right)\right] \\
&\quad + 2\mathrm{E}\left[\left(\tilde{\mu}_b - \bar{Y}_b\right)\left(\mu_b - \mathrm{E}\left[\tilde{\mu}_b\right]\right)\right] + 2\mathrm{E}\left[\left(\bar{Y}_b - \mu_b\right)\left(\mu_b - \mathrm{E}\left[\tilde{\mu}_b\right]\right)\right].
\end{aligned} \tag{39}$$

Since

$$\tilde{\mu}_b|\tilde{F}^{(b)},\mathbf{Y}_b\sim\mathcal{N}\left(\bar{Y}_b,\frac{S_b^2}{n_b}\right)\text{ and }Y_j(\tilde{F}^{(b)})|\tilde{F}^{(b)}\sim\mathcal{N}(\mu_b,\sigma_b^2)\text{ for }j=1,\ldots,n_b,$$

we calculate the cross terms on the right-side of Equation (39) respectively

$$\begin{aligned}
\mathrm{E}\left[\left(\tilde{\mu}_b-\bar{Y}_b\right)\left(\bar{Y}_b-\mu_b\right)\right]&=\mathrm{E}\left\{\left(\bar{Y}_b-\mu_b\right)\mathrm{E}\left[\left(\tilde{\mu}_b-\bar{Y}_b\right)\middle|\tilde{F}^{(b)},\mathbf{Y}_b\right]\right\}=0,\\
\mathrm{E}\left[\left(\tilde{\mu}_b-\bar{Y}_b\right)\left(\mu_b-\mathrm{E}\left[\tilde{\mu}_b\right]\right)\right]&=\mathrm{E}\left\{\left(\mu_b-\mathrm{E}\left[\tilde{\mu}_b\right]\right)\mathrm{E}\left[\left(\tilde{\mu}_b-\bar{Y}_b\right)\middle|\tilde{F}^{(b)},\mathbf{Y}_b\right]\right\}=0,\\
\mathrm{E}\left[\left(\bar{Y}_b-\mu_b\right)\left(\mu_b-\mathrm{E}\left[\tilde{\mu}_b\right]\right)\right]&=\mathrm{E}\left\{\left(\mu_b-\mathrm{E}\left[\tilde{\mu}_b\right]\right)\mathrm{E}\left[\left(\bar{Y}_b-\mu_b\right)\middle|\tilde{F}^{(b)})\right]\right\}=0.
\end{aligned}$$

Thus, we can simplify Equation (39) to

$$\mathrm{Var}\left[\tilde{\mu}(\tilde{F}^{(b)})\right]=\mathrm{E}\left[\left(\tilde{\mu}_b-\bar{Y}_b\right)^2\right]+\mathrm{E}\left[\left(\bar{Y}_b-\mu_b\right)^2\right]+\mathrm{E}\left[\left(\mu_b-\mathrm{E}\left[\tilde{\mu}_b\right]\right)^2\right].\tag{40}$$

Next we calculate each term on the right side of Equation (40) respectively. We can simplify the first term

$$\begin{aligned}
&\mathrm{E}\left[\left(\tilde{\mu}_b-\bar{Y}_b\right)^2\right]=\mathrm{E}\left\{\mathrm{E}\left[\left(\tilde{\mu}_b-\bar{Y}_b\right)^2\middle|\tilde{F}^{(b)},\mathbf{Y}_b\right]\right\}=\mathrm{E}\left[\frac{S_b^2}{n_b}\right]\\
&=\frac{1}{n_b}\mathrm{E}\left\{\mathrm{E}\left[\frac{1}{n_b-1}\sum_{j=1}^{n_b}\left(Y_j(\tilde{F}^{(b)})-\bar{Y}_b\right)^2\middle|\tilde{F}^{(b)}\right]\right\}\\
&=\frac{1}{n_b}\mathrm{E}\left\{\mathrm{E}\left[\frac{1}{n_b-1}\sum_{j=1}^{n_b}\left(Y_j(\tilde{F}^{(b)})-\mu_b+\mu_b-\bar{Y}_b\right)^2\middle|\tilde{F}^{(b)}\right]\right\}\\
&=\frac{1}{n_b(n_b-1)}\mathrm{E}\left\{\mathrm{E}\left[\sum_{j=1}^{n_b}\left(Y_j(\tilde{F}^{(b)})-\mu_b\right)^2+n_b\left(\mu_b-\bar{Y}_b\right)^2\middle|\tilde{F}^{(b)}\right]\right\}\\
&\quad+\frac{2}{n_b(n_b-1)}\mathrm{E}\left\{\mathrm{E}\left[\sum_{j=1}^{n_b}\left(Y_j(\tilde{F}^{(b)})-\mu_b\right)\left(\mu_b-\bar{Y}_b\right)\middle|\tilde{F}^{(b)}\right]\right\}.
\end{aligned}\tag{41}$$

Since $Y_j(\tilde{F}^{(b)})|\tilde{F}^{(b)}\sim\mathcal{N}(\mu_b,\sigma_b^2)$, the second term on the right side of Equation (41) becomes

$$\mathrm{E}\left[n_b\left(\mu_b-\bar{Y}_b\right)^2\right]=n_b\mathrm{E}\left\{\left[\left(\mu_b-\bar{Y}_b\right)^2\middle|\tilde{F}^{(b)}\right]\right\}=\mathrm{E}\left[\sigma_b^2\right],$$

and the third term can be simplified

$$\begin{aligned}
&\mathrm{E}\left\{\mathrm{E}\left[\sum_{j=1}^{n_b}\left(Y_j(\tilde{F}^{(b)})-\mu_b\right)\left(\mu_b-\bar{Y}_b\right)\middle|\tilde{F}^{(b)}\right]\right\}\\
&=\mathrm{E}\left\{\mathrm{E}\left[\sum_{j=1}^{n_b}\left(Y_j(\tilde{F}^{(b)})\mu_b+\bar{Y}_b\mu_b-Y_j(\tilde{F}^{(b)})\bar{Y}_b-\mu_b^2\right)\middle|\tilde{F}^{(b)}\right]\right\}\\
&=\mathrm{E}\left[\mathrm{E}\left(n_b\bar{Y}_b\mu_b+n_b\bar{Y}_b\mu_b-n_b\bar{Y}_b^2-n_b\mu_b^2|\tilde{F}^{(b)}\right)\right]\\
&=\mathrm{E}\left\{\mathrm{E}\left[-n_b\left(\mu_b-\bar{Y}_b\right)^2\middle|\tilde{F}^{(b)}\right]\right\}\\
&=\mathrm{E}\left[-\sigma_b^2\right].
\end{aligned}$$

By submitting these results into Equation (41), we have

$$\mathrm{E}\left[\left(\tilde{\mu}_b - \bar{Y}_b\right)^2\right] = \frac{1}{n_b(n_b-1)}\mathrm{E}\left[n_b\sigma_b^2 + \sigma_b^2 - 2\sigma_b^2\right] = \mathrm{E}\left[\frac{\sigma_b^2}{n_b}\right].$$

Thus, the sum of the first two terms on the right side of Equation (40) becomes

$$\mathrm{E}\left[\left(\tilde{\mu}_b - \bar{Y}_b\right)^2\right] + \mathrm{E}\left[\left(\bar{Y}_b - \mu_b\right)^2\right] = \mathrm{E}\left[\frac{2}{n_b}\sigma_b^2\right],$$

and the third term can be rewrote as

$$\begin{aligned}
&\mathrm{E}\left[\left(\mu_b - \mathrm{E}\left[\tilde{\mu}_b\right]\right)^2\right] = \mathrm{E}\left[\left(\mu_b - \mathrm{E}\left[\mu_b\right] + \mathrm{E}\left[\mu_b\right] - \mathrm{E}\left[\tilde{\mu}_b\right]\right)^2\right] \\
&= \mathrm{E}\left[\left(\mu_b - \mathrm{E}\left[\mu_b\right]\right)^2\right] + \left(\mathrm{E}\left[\mu_b\right] - \mathrm{E}\left[\tilde{\mu}_b\right]\right)^2 + 2\left(\mathrm{E}\left[\mu_b\right] - \mathrm{E}\left[\tilde{\mu}_b\right]\right)\mathrm{E}\left[\mu_b - \mathrm{E}\left[\mu_b\right]\right] \\
&= \mathrm{Var}(\mu_b) + \left(\mathrm{E}\left[\mu_b\right] - \mathrm{E}\left[\tilde{\mu}_b\right]\right)^2 \\
&= \mathrm{Var}(\mu_b). \qquad (42)
\end{aligned}$$

Step (42) holds because $\mathrm{E}[\tilde{\mu}_b] = \mathrm{E}[\mathrm{E}(\tilde{\mu}_b|\tilde{F}^{(b)}, \bar{Y}_b)] = \mathrm{E}[\mathrm{E}(\bar{Y}_b|\tilde{F}^{(b)})] = \mathrm{E}[\mu_b]$. Submitting these results into Equation (40), the total variance of the estimated system response is

$$\mathrm{Var}\left[\tilde{\mu}(\tilde{F}^{(b)})\middle|\,\mathbf{X}_m\right] = \mathrm{Var}(\mu_b|\mathbf{X}_m) + \mathrm{E}\left[\frac{2}{n_b}\sigma_b^2\middle|\,\mathbf{X}_m\right]. \qquad (43)$$

The first term on the right side of Equation (43) measures the impact of input uncertainty. The second term which is the expected simulation estimation uncertainty weighted by $p(F|\mathbf{X}_m)$ measures the impact from the simulation uncertainty. □

Appendix D: Sensitivity Analysis of Hyper-parameters for $\boldsymbol{\theta}_\alpha$

We use examples listed in Table 1 with sample size $m = 50$ to study the sensitivity to the values of hyper-parameters $\boldsymbol{\theta}_\alpha$. DPM with appropriate kernel densities are used for different examples. That means DPM with Gamma kernel used for Example 1 and 2, DPM with Gaussian kernel used for Examples 3, and DPM with Beta kernel used for Example 4. Table 4 records 95% symmetric CIs of KS and AD distances obtained from 1000 macro-replications. The results indicate that the values of hyper-parameters $\boldsymbol{\theta}_\alpha$ have an insignificant impact on the input model estimation, where Gamma$(2, 4)$ prior was used in Escobar and West (1995) and the discrete Gamma$(1, 1)$ prior was used in Wang and Dunson (2011). The choice of hyper-parameters does not have significant impact on the density estimation accuracy.

Table 4 KS and AD distances for Examples 1–4 with different hyper-parameters $\boldsymbol{\theta}_\alpha$

$m = 50$		Example 1	Example 2	Example 3	Example 4
Gamma(0.5, 0.5)	D_m	0.106±0.002	0.073±0.001	0.075±0.001	0.070±0.001
	A_m	11.870±0.175	7.594±0.097	6.365±0.096	8.808±0.097
Gamma(1, 1)	D_m	0.102±0.002	0.071±0.001	0.072±0.001	0.068±0.001
	A_m	11.278±0.158	7.203±0.088	6.083±0.093	8.253±0.092
Gamma(4, 4)	D_m	0.104±0.002	0.074±0.001	0.075±0.001	0.069±0.001
	A_m	11.495±0.166	7.787±0.104	6.484±0.098	8.490±0.095
Gamma(2, 4)	D_m	0.105±0.002	0.072±0.001	0.073±0.001	0.068±0.001
	A_m	11.762±0.174	7.419±0.092	6.207±0.094	8.337±0.094

Table 5 Results of cross validation for the density selection

	DPM Gamma	Empirical Distribution	KDE	Parametric
RM1	-218.731	NA	-226.928	-393.239
RM2	-233.971	NA	-270.577	-605.704

Appendix E: Studying Input Model Performance by Using Real Demand Data

Except the simulated data used in Section 4.1, we also test the performance of our nonparametric input models by using the demand data of two representative raw materials collected from a real bio-pharmaceutical inventory system. The sample sizes are 101 and 142 respectively. Since the underlying true distributions are unknown, the cross validation is applied for the density selection; See more detailed description in Lian (2009). We perform a 5-folds cross validation. Table 5 records the average log-likelihoods obtained by using different approaches. Specifically, we randomly divide all the data into 5 sets, select one set for validation and use the remaining sets as training data. For each combination of training and validation data sets, we first fit the input model by using the training data, apply it to the validation data and calculate the log-likelihood. After that, we record the average log-likelihood obtained from all combinations of training and validation data sets.

Since the demand data have support on $\Re^+$, we use DPM with Gamma kernel density. The distribution family for the parametric approach is selected based on the KS test statistics by using @Risk since both this criteria and the likelihood are related to the overall fitting performance of input model. In addition, since the empirical distribution only has the information at the data points and it does not return a density estimate, we skip it.

Since the posterior predictive distribution is recommended for the model selection (Gelman et al. 2004), for DPM, the likelihood is calculated based on the posterior predictive distribution: $f\big(\mathbf{X}_V^{(i)}|\mathbf{X}_T^{(i)}\big) = \int f\big(\mathbf{X}_V^{(i)}|F\big)dP\big(F|\mathbf{X}_T^{(i)}\big)$, where $\mathbf{X}_T^{(i)}$ and $\mathbf{X}_V^{(i)}$ denote the ith combination

of training and validation data with $i = 1, 2, \ldots, 5$. Then, we record the average log-likelihood $\sum_{i=1}^{5} \log\left[f\left(\mathbf{X}_V^{(i)}|\mathbf{X}_T^{(i)}\right)\right]/5$. For the frequentist KDE and parametric approaches, we first find the fitted input density based on the training set, denoted by $\widehat{f}(\cdot|\mathbf{X}_T^{(i)})$, apply it to the validation data and calculate the average log-likelihood $\sum_{i=1}^{5} \log\left[\widehat{f}\left(\mathbf{X}_V^{(i)}|\mathbf{X}_T^{(i)}\right)\right]/5$. Table 5 demonstrates that DPM with Gamma kernel maximizes the average log-likelihood and provides the best fit to the real RM demand data.

Appendix F: Empirical Study for an Inventory Example

In this section, we use a RM inventory example to study the performance of our Bayesian framework. A daily review (R, Q) ordering policy is applied for the inventory control. At the beginning of each day, if the inventory position drops to and below the reorder point $R = 150$, we place an order with size $Q = 300$ that arrives in a fixed lead time with $L = 1$. In each day, the arrivals of move orders follow a Poisson process with rate 3 and the size of each move order follows the Log-normal mixture distribution $0.3L(-0.005, 0.1) + 0.4L(0.378, 0.2) + 0.3L(0.654, 0.3)$ to mimic the situations that we could have various latent sources of uncertainty from production lines. Thus, the accumulated move orders in each day follow the compound Poisson distribution. Notice that compared to the $M/M/1$ queue used in Section 4.2, this input model is more complex. Without any prior information of F^c, it is challenging for existing input modeling approaches to capture the important properties in the real-world data. We are interested in the steady-state expected inventory level and two most commonly used service levels in the inventory control: the type-I service level defined as the probability of no stockout per order cycle and the type-II service level defined as the fraction of move orders that can be satisfied immediately from stock on hand. Since the performance of our approach for type-I and type-II service levels are similar, we only present the results for type-II service level.

A side experiment driven by the underlying input model F^c is used to estimate the true system performance μ^c. We start with the empty system with the warmup length equal to 100 days, the run length equal to 500 days and the number of replications equal to 10^6. We record 95% symmetric CIs of the expected inventory level 90.261 ± 0.006 and the type-II service level 0.82025 ± 0.00009.

When we evaluate the performance of our framework, the simulation is used to estimate the steady-state system response for the inventory system. In each simulation run, we start with an empty system, set the warmup length equal to 100 days and the runlength equal to 50 days. To create the situation where each simulation run could be computationally expensive, we choose a short runlength on purpose.

We systematically examine the effects of m and n. The results of the expected inventory level and the type-II service level with $m = 50, 100, 1000$ and $n = 10, 50, 500$ are shown in Tables 6–7.

All results in Tables 6-7 are based on 1000 macro-replications. In each macro-replication, we first draw m samples from F^c to mimic the procedure collecting m "real-world" data. As m and n increase, the width of CrI quantifying the overall uncertainty of the system performance estimation decreases and the deviation $|\mathrm{E}[U|\mathbf{X}_m, \mathcal{Y}_n] - \mu^c|$ also decreases. These behaviors match well with the asymptotic properties described in Section 3.5.2. Thus, the performance of our approach is robust to complex input model. Since there is no closed form of the system response as a function of the input model and it is also computationally expensive to precisely estimate the true mean response at each posterior sample of input model, we do not provide the results of the PC for this example.

Table 6 Results of the expected inventory level when $m = 50, 100, 1000$ and $n = 10, 50, 500$.

	Mean of |CrI|/2	SD of |CrI|/2	Mean of |err|	SD of |err|	$\{\widehat{\sigma}_I/\widehat{\sigma}_S\}$
$m = 50, n = 10$	22.643	1.929	12.045	7.912	1.498
$m = 50, n = 50$	19.815	1.523	9.744	7.793	3.019
$m = 50, n = 500$	18.162	1.555	9.397	8.296	9.334
$m = 100, n = 10$	18.672	1.193	7.627	5.883	1.162
$m = 100, n = 50$	14.026	0.886	7.148	5.852	2.184
$m = 100, n = 500$	12.574	0.818	7.407	5.297	6.508
$m = 1000, n = 10$	14.082	0.492	2.229	1.65	0.745
$m = 1000, n = 50$	6.86	0.261	2.462	1.74	0.859
$m = 1000, n = 500$	3.571	0.143	2.375	1.774	1.691

Table 7 Results of the type-II service level when $m = 50, 100, 1000$ and $n = 10, 50, 500$.

	Mean of |CrI|/2	SD of |CrI|/2	Mean of |err|	SD of |err|	$\{\widehat{\sigma}_I/\widehat{\sigma}_S\}$
$m = 50, n = 10$	0.171	0.018	0.041	0.026	1.466
$m = 50, n = 50$	0.143	0.014	0.033	0.024	2.937
$m = 50, n = 500$	0.135	0.015	0.029	0.03	9.084
$m = 100, n = 10$	0.145	0.012	0.028	0.02	1.159
$m = 100, n = 50$	0.117	0.01	0.027	0.021	2.161
$m = 100, n = 500$	0.099	0.008	0.025	0.02	6.476
$m = 1000, n = 10$	0.092	0.004	0.009	0.006	0.745
$m = 1000, n = 50$	0.054	0.002	0.008	0.006	0.868
$m = 1000, n = 500$	0.028	0.001	0.006	0.006	1.717

Acknowledgments

The authors are grateful for helpful and constructive comments from Barry L. Nelson.

References

Akcay A, Biller B, Tayur S (2011) Improved inventory targets in the presence of limited historical demand data. *Manufacturing and Service Operations Management* 13(3):297–309.

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

Barton RR (2007) Presenting a more complete characterization of uncertainty: Can it be done? *Proceedings of the 2007 INFORMS Simulation Society Research Workshop* (Fontainebleau: INFORMS Simulation Society).

Barton RR (2012) Tutorial: Input uncertainty in output analysis. C Laroque, J Himmelspach, R Pasupathy, O Rose, and AM Uhrmacher, ed., *Proceedings of the 2012 Winter Simulation Conference*, 67–78 (Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.).

Barton RR, Nelson BL, Xie W (2014) Quantifying input uncertainty via simulation confidence intervals. *Informs Journal on Computing* 26:74–87.

Barton RR, Schruben LW (1993) Uniform and bootstrap resampling of input distributions. *Proceedings of the 1993 Winter Simulation Conference*, 503–508 (Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.).

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

Billingsley P (1999) *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics (New York: John Wiley & Sons Inc.), second edition, ISBN 0-471-19745-9, a Wiley-Interscience Publication.

Bishop CM (2006) *Pattern Recognition and Machine Learning* (New York: Springer).

Cheng RCH, Currie CSM (2003) Prior and candidate models in the bayesian analysis of finite mixtures. *Simulation Conference, 2003. Proceedings of the 2003 Winter*, volume 1, 392–398 Vol.1.

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

Escobar MD (1994) Estimating normal means with a Dirichiet process prior. *Journal of the American Statistical Association* 89(425):268–277.

Escobar MD, West M (1995) Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* 90(430):577–588.

Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* 1(196):209–230.

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

Ghosal S, Ghosh JK, Ramamoorthi RV (1999) Posterior consistency of Dirichlet mixtures in density estimation. *Annals of Statistics* 27(1):143–158.

Ghosal S, Roy A, Tang Y (2008) Posterior consistency of Dirichlet mixtures of beta densities in estimating positive false discovery rates. *IMS Collections: Beyond parametrics in interdisciplinary research: Festschrift in honor of Professor Prabnab K. Sen* 1:105–115.

Ghosh JK, Ramamoorthi RV (2003) *Bayesian Nonparametrics* (New York: Springer–Verlag).

Görür D, Rasmussen CE (2010) Dirichlet process Gaussian mixture models: Choice of the base distribution. *Journal of Computer Science and Technology* 25(4):653–664.

Hanson TE (2006) Modeling censored lifetime data using a mixture of gammas baseline. *Bayesian Anal.* 1(3):575–594, URL `http://dx.doi.org/10.1214/06-BA119`.

Hjort NL, Holmes C, P M, Walker SG (2011) *Bayesian Nonparametrics* (New York: Cambridge University Press).

Kass RE, Raftery AE (1995) Bayes factors. *Journal of the American Statistical Association* 90(430):773–795, ISSN 01621459, URL `http://www.jstor.org/stable/2291091`.

Kottas A (2006) Dirichlet process mixtures of beta distributions, with applications to density and intensity estimation. Technical report, University of California, Santa Cruz.

Lam H (2016) Advanced tutorial: Input uncertainty and robust analysis in stochastic simulation. Roeder TMK, Frazier PI, Szechtman R, Zhou E, Huschka T, Chick SE, eds., *Proceedings of the 2016 Winter Simulation Conference* (Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.).

Laurent B, Massart P (2000) Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics* 28(5):1302–1338.

Law AM (2015) *Simulation Modeling and Analysis* (New York: Wiley), 5 edition.

Lian H (2009) Cross-validation for comparing multiple density estimation procedures. *Statistics & Probability Letters* 79(1):112–115.

Lo AY (1984) On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics* 12(1):351–357.

Ma Y (2011) *Risk Management in Biopharmaceutical Supply Chains*. Ph.D. thesis, University of California, Berkeley.

MacEachern S, Muller P (2000) *Efficient MCMC Schemes for Robust Model Extensions Using Encompassing Dirichlet Process Mixture Models*, 295–315 (New York, NY: Springer New York), ISBN 978-1-4612-1306-2.

Maceachern SN (1994) Estimating normal means with a conjugate style dirichlet process prior. *Communications in Statistics - Simulation and Computation* 23(3):727–741.

MacEachern SN (1998) *Computational Methods for Mixture of Dirichlet Process Models*, 23–43 (New York, NY: Springer New York), ISBN 978-1-4612-1732-9.

Massart P (2007) *Concentration inequalities and model selection* (New York: Springer).

Neal RM (2000) Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* 9(2):249–265.

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

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

Petrone S, Wasserman L (2002) Consistency of bernstein polynomial posteriors. *Journal of Royal Statistical Society: Series B* 64:79–100.

Rousseau J (2010) Rates of convergence for the posterior distributions of mixtures of betas and adaptive nonparametric estimation of the density. *The Annals of Statistics* 38:146–180.

Sheather SJ, Jones MC (1991) A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B* 53:683–690.

Song E, Nelson BL, Pegden CD (2014) Advanced tutorial: Input uncertainty quantification. Tolk A, Diallo SY, Ryzhov IO, Yilmaz L, Buckley S, Miller JA, eds., *Proceedings of the 2014 Winter Simulation Conference* (Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.).

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002) Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B* 64(4):583–639, URL `http://EconPapers.repec.org/RePEc:bla:jorssb:v:64:y:2002:i:4:p:583-639`.

Tokdar ST (2006) Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā: The Indian Journal of Statistics* 68:90–110.

Wagner SM, Bode C, Koziol P (2009) Supplier default dependencies: Empirical evidence from the automotive industry. *European Journal of Operational Research* 199:150–161.

Wang L, Dunson DB (2011) Fast Bayesian inference in dirichlet process mixture models. *Journal of Computational and Graphical Statistics* 20(1):196–216.

West M (1990) Bayesian kernel density estimation. Technical report, Duke University, Institute of Statistics and Decision Sciences.

Wiper I (2001) Mixtures of gamma distributions with applications. *Journal of Computational & Graphical Statistics* 10(3):440–454.

Wu Y, Ghosal S (2008) Kullback leibler property of kernel mixture priors in bayesian density estimation. *Electronic Journal of Statistics* 2:298–331.

Xie W, Nelson BL, Barton RR (2014) A bayesian framework for quantifying uncertainty in stochastic simulation. *Operations Research* 62(6):1439–1452.

Yi Y, Xie W (2017) A design of experiments for quantifying the impact of input uncertainty in stochastic simulation. *ACM Transactions on Modeling and Computer Simulation* Accepted.

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

Zouaoui F, Wilson JR (2004) Accounting for input-model and input-parameter uncertainties in simulation. *IIE Transactions* 36(11):1135–1151.