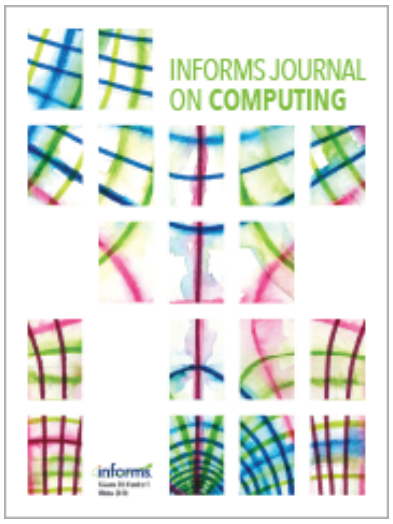


INFORMS Journal on Computing

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

Reducing Simulation Input-Model Risk via Input Model Averaging

Barry L. Nelson, Alan T. K. Wan, Guohua Zou, Xinyu Zhang, Xi Jiang

To cite this article:
Barry L. Nelson, Alan T. K. Wan, Guohua Zou, Xinyu Zhang, Xi Jiang (2020) Reducing Simulation Input-Model Risk via Input Model Averaging. INFORMS Journal on Computing

Published online in Articles in Advance 06 Oct 2020

. https://doi.org/10.1287/ijoc.2020.0994

http://pubsonline.informs.org/journal/ijoc

INFORMS JOURNAL ON COMPUTING
Articles in Advance, pp. 1–13
ISSN 1091-9856 (print), ISSN 1526-5528 (online)

Reducing Simulation Input-Model Risk via Input Model Averaging

Barry L. Nelson,[a] Alan T. K. Wan,[b] Guohua Zou,[c] Xinyu Zhang,[d,e,*] Xi Jiang[a]

[a] Northwestern University, Evanston, Illinois 60208-3119; [b] City University of Hong Kong, Kowloon, Hong Kong; [c] Capital Normal University, Beijing 100048, China; [d] University of Science and Technology of China, Hefei 230052, China; [e] Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

*Corresponding author

Contact: nelsonb@northwestern.edu, https://orcid.org/0000-0002-1325-2624 (BLN); alan.wan@cityu.edu.hk (ATKW); ghzou@amss.ac.cn (GZ); xinyu@amss.ac.cn (XZ); xijiang2020@u.northwestern.edu (XJ)

Received: January 22, 2017
Revised: December 31, 2017; February 28, 2019; May 11, 2020; May 21, 2020; June 11, 2020
Accepted: June 12, 2020
Published Online in Articles in Advance: October 6, 2020

https://doi.org/10.1287/ijoc.2020.0994

Abstract. Input uncertainty is an aspect of simulation model risk that arises when the driving input distributions are derived or "fit" to real-world, historical data. Although there has been significant progress on quantifying and hedging against input uncertainty, there has been no direct attempt to reduce it via better input modeling. The meaning of "better" depends on the context and the objective: Our context is when (a) there are one or more families of parametric distributions that are plausible choices; (b) the real-world historical data are not expected to *perfectly* conform to any of them; and (c) our primary goal is to obtain higher-fidelity simulation *output* rather than to discover the "true" distribution. In this paper, we show that frequentist model averaging can be an effective way to create input models that better represent the true, unknown input distribution, thereby reducing model risk. Input model averaging builds from standard input modeling practice, is not computationally burdensome, requires no change in how the simulation is executed nor any follow-up experiments, and is available on the Comprehensive R Archive Network (CRAN). We provide theoretical and empirical support for our approach.

History: Accepted by Bruno Tuffin, Area Editor for Simulation.
Funding: B. L. Nelson's work was partially supported by the National Science Foundation [Grant CMMI-1634982]. A. T. K. Wan's work was supported by the City University of Hong Kong [Grant 7004985], the Hong Kong Research Grants Council [Grant CityU 11500419], and the National Natural Science Foundation of China [Grant 71973116]. G. Zou's work was supported by the Ministry of Science and Technology of China [Grant 2016YFB0502301] and the National Natural Science Foundation of China [Grants 11971323 and 11529101]. X. Zhang's work was supported by the National Natural Science Foundation of China [Grants 71925007, 71522004 and 71631008] and Youth Innovation Promotion Association of the Chinese Academy of Sciences.
Supplemental Material: Data and the online appendices are available at https://doi.org/10.1287/ijoc.2020.0994.

Keywords: input modeling • stochastic simulation • input uncertainty

1. Introduction

At a high level, stochastic simulations consist of inputs and logic. The inputs are the basic sources of uncertainty that defy further explanation; they are represented by fully specified probability models (e.g., exponential distribution with rate $\lambda = 7.2$). The logic consists of rules and algorithms that transform realizations of the inputs into sample paths of system performance (e.g., waiting times in queues); estimating system properties from these sample paths or "outputs" is the reason a simulation experiment is performed. The fidelity of the outputs in representing the performance of a real or conceptual system clearly depends—often in a very complex way—on the fidelity of the input models. In this paper, we consider input models that are tuned or "fit" to samples of real-world, historical data. We refer to this activity as *input modeling*, and we propose a better way to do it.

Beyond the availability of good software, methods used for input modeling in the stochastic simulation practice community have not advanced much in the last 30 years.[1] Here is the recipe found in many textbooks (e.g., Law and Kelton 1991, Banks et al. 2010) for fitting a marginal distribution F to describe an independent and identically distributed (i.i.d.) input process:

1. Let $x_1, x_2, \ldots, x_N$ be an i.i.d. sample from some unknown input distribution F^c, with "c" denoting "correct" or "true" distribution.

2. Fit $q \geq 1$ candidate parametric distributions $\mathcal{F} = \{F_1, F_2, \ldots, F_q\}$ using methods such as maximum likelihood estimation (MLE), least squares, or moment matching. This yields a set of fitted distributions, say, $\widehat{\mathcal{F}} = \{\widehat{F}_1, \widehat{F}_2, \ldots, \widehat{F}_q\}$. The number of choices in current software ranges from $q = 10$ to 40.

3. Rank the choices using one or more summary measures of fit. Standard measures are hypothesis-test statistics such as chi-squared, Kolmogorov–Smirnov

and Anderson–Darling, and likelihood-based statistics such as AIC and BIC.

4. From among the top choices, evaluate the fit, for example, via p-values of the hypothesis tests or graphically via Q-Q plots or other tools.

5. Select $\widehat{F} = \text{Best Choice}\{\widehat{F}_1, \widehat{F}_2, \ldots, \widehat{F}_q\}$. Alternatively, use the (possibly smoothed) empirical distribution of $x_1, x_2, \ldots, x_N$ if nothing fits well.

Although this recipe can and should be made smarter, for instance, by using the physical basis of the real input process to suggest an appropriate family of distributions, in practice, step 5 is often automated by selecting the distribution with, say, the minimum AIC statistic for each input process, bypassing step 4. This approach is understandable because it is not obvious to simulation practitioners either how to do better or how much the choice actually matters. Our proposal adopts step 2, but rather than selecting one element of $\mathcal{F}$, it instead creates an "input model average" that often leads to a better input model.

Our work is motivated by the current interest in simulation model risk due to input uncertainty, which is the uncertainty resulting from having only a finite sample N of real-world data from which to fit $\widehat{F}$. However, the input-uncertainty literature has emphasized either quantifying the variability in the simulation output due to the sampling variability in $\widehat{F}$ or selecting a defensive $\widehat{F}$, which means a distribution that is plausible with respect to the given data but leads to the worst-case (maximum or minimum) simulation output performance (see, e.g., Lam 2016). Our objective is to reduce input uncertainty through our choice of $\widehat{F}$ via a rethinking about how the input models are created. Our work is heavily motivated by recent advances in the statistics literature on using model averaging within the frequentist paradigm to improve parameter estimation efficiency and forecast accuracy (e.g., Hjort and Claeskens 2003, Hansen 2007, Wan et al. 2010, Liang et al. 2011).

What do we want in an input modeling solution?

- It should work within the framework of current input-modeling software and, in particular, step 2, where we have a collection of candidate distributions and impose only a modest additional computational burden.
- It should not require any change in how we actually execute the simulation, other than generating inputs from a different $\widehat{F}$. Thus, input modeling remains an upfront step in the simulation experiment (input uncertainty quantification, on the other hand, often requires additional follow-up experiments).
- It should improve simulation *output* fidelity when the true distribution is *not* in $\mathcal{F}$—so no single choice can be fully correct—but also tends to favor a single candidate model when it is close to F^c. This is consistent with the "view through the queue" criterion coined by Whitt (1981), which evaluates an input approximation by how well it reproduces the desired output, rather than whether it discovers the true input.

More pointedly, our model averaging approach is *not* a better way to discover the "true" real-world distribution when it is a member of the candidate set $\mathcal{F}$, either individually or as a mixture. In fact, our asymptotic analysis specifically assumes that $F^c \notin \mathcal{F}$ and shows that, under some assumptions on the candidate set of distributions $\mathcal{F}$, our model-average distribution gets as close as possible to the real-world distribution using only the component distributions in $\mathcal{F}$. Thus, model averaging is not generally consistent for F^c; however, if we include the empirical distribution (ED) as a candidate, then the model average places all weight on the ED as the sample size $N \to \infty$. Further, under very weak assumptions, the empirically optimal model-average distribution exists and is easily found.

In the end, we will recommend the following: Reduce the size of the candidate set $\mathcal{F}$ (if large) by using prior knowledge of the input process physics or by screening out poor choices using something like AIC or BIC; always include the ED in $\mathcal{F}$; and then do model averaging. However, model averaging can be applied in a completely automated fashion to a large candidate set, and the ED need not be included (say if a continuous $\widehat{F}$ is desired).

The paper is organized as follows. We describe the problem and context more fully in Section 2 and our input model averaging method in Section 3. An empirical evaluation follows in Section 4. Proofs of some of the results are found in Online Appendices A–B.

2. Background and Examples

In this paper, we focus on univariate input models, but the method extends naturally to random vectors. Generically, X and Y denote input and output random variables, respectively, and x and y are realized values. We use F_Y to refer to the (typically unknown) distribution of Y.

The following examples will be used to evaluate our methods; they were chosen because they mimic three distinct classes of problems found in simulation studies and because we expect that different aspects of the inputs X (e.g., mean, variance, tail behavior) will affect the fidelity of their outputs Y. That is, even though the examples themselves are simple, they manifest complex and differing input-to-output behavior. The examples are important because we rely solely on empirical evaluation to establish the potential reduction in input uncertainty.

2.1. Stochastic Activity Network (SAN)

Stochastic activity networks are used in project planning when there is interest in the time to complete the project; an early paper on simulating such networks is Burt and Garman (1971). A realistic problem might have several hundred activities, constrained resources,

and so on, but, as an illustration, we consider one with five activities where the time to complete the project is

$$Y = \max\{X_1 + X_4, X_1 + X_3 + X_5, X_2 + X_5\}. \quad (1)$$

See Figure 1. Thus, simulation of the SAN requires five input distributions for $X_1, X_2, \ldots, X_5$. Properties of Y that are of interest include $\mathrm{E}(Y)$, $\Pr\{Y > c\}$, $F_Y^{-1}(p)$, or the entire distribution. The natural experiment design for the simulation is to make R replications yielding i.i.d. outputs $Y_1, Y_2, \ldots, Y_R$. Each replication requires random activity-time inputs $X_1, X_2, \ldots, X_5$. Because activity times are summed, path durations tend to be normally distributed for realistically large projects, but, for this small example, the specific distributions of the X_i should matter.

2.2. *GI/G/1* Queue

The $GI/G/1$ queue has a renewal arrival process of customers, some nonnegative service-time distribution, and a single server (see, e.g., Gross et al. 2008). Let Y_i be the delay in queue of the ith arriving customer when the system starts empty. Then,

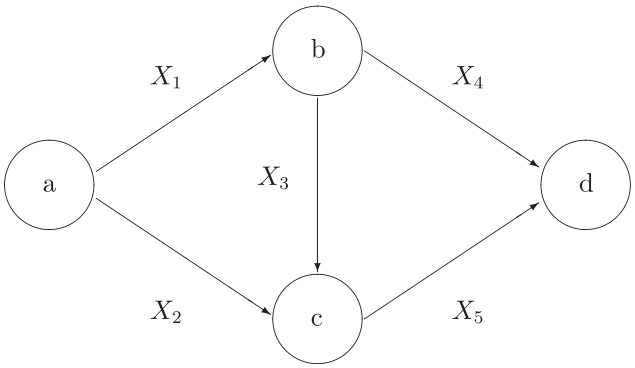
$$Y_i = \max\{0, Y_{i-1} + X_{2,i-1} - X_{1,i}\}, \; i = 1, 2, \ldots, \quad (2)$$

with $Y_0 = X_{2,0} = 0$. There are two input distributions, the interarrival-time distribution of X_1 and the service-time distribution of X_2; successive interarrival times and service times are individually and jointly independent.

Under certain conditions, it can be shown that $Y_i \xrightarrow{\mathcal{D}} Y$ as $i \to \infty$, and properties of Y, a specific Y_i, or the set $\{Y_i, i \le T\}$ for some stopping time T are of interest. Thus, the experiment design could be a single, long replication or multiple shorter ones, and the number of inputs $X_{1,i}$ and $X_{2,i}$ needed may be fixed or random.

The $\mathrm{E}(Y)$ is tractable if the interarrival times are exponentially distributed and it depends only on the mean and variance of the service times X_2; the entire distribution of Y is tractable if the service time is also exponential. In general, the distributions of Y_i and Y are not known. In our evaluation, we focus on the distribution of Y_5, the wait of the fifth arriving customer, since the effect of the service-time distribution beyond its mean and variance should not yet have washed out.

Figure 1. A Small Stochastic Activity Network



2.3. Highly Reliable System (HRS)

Systems that are repairable or have significant redundancy are designed to be highly reliable, meaning that system failure is rare. Let Y be the time of system failure. The following algorithm mimics an HRS for which a failure is avoided if a backup component is repaired (time to repair X_1) before the active component fails (time to failure X_2); it does not actually model such a system but allows us to control the rarity of failure through the distributions of X_1 and X_2:

1. $Y = 0$; $i = 1$
2. until $X_{2,i} < X_{1,i}$ do
   $Y = Y + X_{2,i}$
   $i = i + 1$
   loop
3. return $Y = Y + X_{i,1}$.

If $\mathrm{E}(X_1) \ll \mathrm{E}(X_2)$, then the system will be highly reliable—just how reliable is determined by properties of Y, such as its mean or a tail probability. In our example, $X_{i,1}$ and $X_{i,2}$ are individually and jointly independent.

2.4. Input Uncertainty

To present our method, we focus on a single, univariate input distribution F^c for which we have an i.i.d. sample $x_1, x_2, \ldots, x_N$ of real-world data. Because it is a real-world process, there is no expectation that F^c is a member of any standard family of parametric distributions, including those in our set $\mathcal{F}$, although some may be close.

The distribution of our generic output Y depends upon the choice of input distribution F; thus, we write

$$Y \sim F_Y(y \mid F).$$

Based on the input data, we fit a distribution denoted by $\hat{F}$; thus, the simulation generates observations of

$$\hat{Y} \sim F_Y(y \mid \hat{F}).$$

Ideally, $F_Y(y \mid \hat{F}) = F_Y(y \mid F^c)$, but, in practice, we will be satisfied if the distributions are close in some relevant sense (e.g., have nearly the same mean). Notice that what matters is the implied output distribution; the input distribution F^c itself is of less interest. We let $Y^c \sim F_Y(y \mid F^c)$ denote the ideal output.

Research on *input uncertainty* addresses the problem that

$$F_Y(y \mid \hat{F}) \neq F_Y(y \mid F^c)$$

often by focusing on the error in using $\hat{Y}$ as an estimator of $\mathrm{E}(Y^c)$. See, for instance, the surveys in Barton (2012), Lam (2016), and Song et al. (2014).

One reasonable objective is to form a confidence interval or a Bayesian credible interval for $E(Y^c)$ that accounts for error in using $\hat{F}$ as an estimator of F^c, as well as the stochastic error from observing the simulated output $\hat{Y}$ rather than $E(\hat{Y})$. There has been significant success in attacking this and related problems, including Cheng and Holland (1997), Cheng and Holland (1998), Chick (2001), Zouaoui and Wilson (2004), Ankenman and Nelson (2012), Barton et al. (2013), Corlu and Biller (2013), Fan et al. (2013), Xie et al. (2014), Song and Nelson (2015), Ghosh and Lam (2015), Song et al. (2015), Zhou and Xie (2015), Glynn and Lam (2018), and Lam and Qian (2018), to name a few. Notice that none of these papers attempt to *reduce* input uncertainty; instead, they try to quantify it or hedge against it.

Unfortunately, experience has shown that the added error due to input uncertainty can be substantial, and sometimes overwhelming, even when we have a significant quantity of real-world data. Therefore, in this paper, we look to reduce the input-uncertainty error by our choice of $\hat{F}$, a reduction that would then be reflected in reduced measurements of it using the methods described in the aforementioned papers. We are not attempting to create a defensive choice $\widehat{F}_{\text{worst}}$, and, in fact, our approach would be an impediment to such robust analysis (see Lam 2016 for an excellent survey of these methods).

Reducing the effect of input-model uncertainty on the simulation output is challenging for obvious reasons. The standard families of distributions used in simulations are often supported by process physics; they are flexible, meaning that they can assume a variety of shapes; and they are accompanied by provably efficient parameter-estimation methods, such as MLE. Improving the standard approach universally would be impossible, but we will demonstrate empirically that substantial improvements are possible in some situations, especially when the real-world input data do not perfectly conform to any known parametric distribution, as is frequently the case in practice. For completeness, we also evaluate our model-average distribution $\widehat{F}$ against F^c, which we can do because we will create the input data.

3. Input Model Averaging

To motivate the method that follows, recall that we have a set $\mathcal{F}$ of q candidate parametric distributions for F^c; for instance, $\mathcal{F}$ could contain the following:

1. exponential: $f_1(x) = \theta e^{-x\theta}$, $x \in [0, \infty)$,
2. normal: $f_2(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-(x-\mu)^2/(2\sigma^2)}$, $x \in \mathbb{R}$,
3. shifted gamma: $f_3(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} (x - \xi)^{(\alpha-1)} e^{-(x-\xi)\beta}$, $x \in (\xi, \infty)$,

where $f_m(x), m = 1, 2, 3$ are density functions for $\mathcal{F} = \{F_1, F_2, F_3\}$, and $\theta, \sigma, \beta, \alpha > 0$ and $\mu, \xi \in \mathbb{R}$ are unknown parameters.

Let $\widehat{F}_m(x)$ and $\widehat{f}_m(x)$ be the estimators of $F^c(x)$ and $f^c(x)$ under the mth candidate distribution, and let $\mathbf{w} = (w_1, w_2, \ldots, w_q)^T$ be a weight vector belonging to the set $\mathcal{W} = \{\mathbf{w} \in [0,1]^q : \sum_{m=1}^q w_m = 1\}$. The *model-average estimator* of $F^c(x)$ is

$$\widehat{F}(x, \mathbf{w}) = \sum_{m=1}^{q} w_m \widehat{F}_m(x), \tag{3}$$

and, taking its derivative with respect to x, we have the model-average estimator of $f^c(x)$:

$$\widehat{f}(x, \mathbf{w}) = \sum_{m=1}^{q} w_m \widehat{f}_m(x). \tag{4}$$

Clearly, $\widehat{F}(x, \mathbf{w})$ includes each of the individual candidate distributions as a special case of $\mathbf{w}$, but it increases flexibility by allowing averages. A good choice of $\mathbf{w}$ is one that makes $\widehat{F}(x, \mathbf{w})$ close to $F^c(x)$ in a comprehensive way, which we will define precisely. Of course, $F^c(x)$ is unknown, but the ED with cumulative distribution function (cdf)

$$\bar{F}(x) = N^{-1} \sum_{i=1}^{N} I(x_i \le x)$$

is unbiased and consistent for it, and so we use $\bar{F}$ as a stand-in for F^c in fitting. Finally, to guard against overfitting, we use cross-validation (CV) with $\bar{F}(x)$ to select $\mathbf{w}$; we describe the CV method in the next section. Given the CV-estimated weight $\widehat{\mathbf{w}}$, variate generation is easy: Each time a value of X is needed, generate integer $M \sim \widehat{\mathbf{w}}$ to select the distribution, and then generate $X \sim \widehat{F}_M$.

Remark 1. Averaging distributions that are as dissimilar as exponential, normal and shifted-gamma may not seem sensible. However, practitioners often use software that fits a long list of distributions, and, as we will show, we can easily find the empirically optimal model average even for such heterogeneous cases.

Remark 2. The model-average distribution $\widehat{F}(x, \mathbf{w})$ is clearly a mixture distribution, but it is different in spirit from mixing a finite number of distributions from a *common* family, which is a well-known distribution fitting approach (McLachlan and Peel 2004). We can, in fact, exploit finite mixtures of a common distribution by including such models in the candidate set $\mathcal{F}$, provided that we have a method for fitting them.

3.1. Cross-Validation for Input Model Averaging

Let $\mathbf{x}_N = (x_1, x_2, \ldots, x_N)$ be the real-world data, which we model as i.i.d. copies of the random variable $X \sim F^c$. Here we develop a "frequentist model averaging" approach to estimate $F^c(x)$ by $F(x, \mathbf{w})$ using J-fold CV to tune $\mathbf{w}$ to $\mathbf{x}_N$; it is in the spirit of the

Jackknife model average (JMA) of Hansen and Racine (2012), developed originally for improving the efficiency of estimators in a heteroscedastic linear regression model. Hansen and Racine (2012) proved that the JMA estimator of the regression coefficients has the smallest asymptotic expected squared errors among a large class of linear estimators, including the least squares, ridge, Nadaraya–Watson, and spline estimators. They also showed that the JMA estimator frequently outperforms the AIC and BIC model selection estimators, as well as Hansen (2007)'s Mallows model-average estimators in finite samples. Zhang et al. (2013) showed that the merits of the JMA estimator carry over to models that admit a lagged dependent variable as a regressor and a nondiagonal error covariance structure.

To implement the JMA scheme for input modeling in stochastic simulation, we partition the data set $\mathbf{x}_N$ into J groups, such that, for each group, we have $S = N/J$ observations. For the jth group, the observations are labeled as $x_{(j-1)S+1}, \ldots, x_{jS}$, where $j = 1, 2, \ldots, J$. Let $\widetilde{F}_m^{(-j)}(x)$ be the estimator (e.g., via MLE) of $F^c(x)$ with the observations of the jth group *removed* from the data set for the mth candidate distribution. Correspondingly, the model-average estimator with the jth group removed is

$$\widetilde{F}^{(-j)}(x, \mathbf{w}) = \sum_{m=1}^{q} w_m \widetilde{F}_m^{(-j)}(x).$$

The ED estimator of $F^c(x)$ using *only* the jth group is

$$\bar{F}_{(j)}(x) = S^{-1} \sum_{s=1}^{S} I\left(x_{(j-1)S+s} \le x\right), \tag{5}$$

and it is well known that $\mathrm{E}(\bar{F}_{(j)}(x)) = F^c(x)$. Our J-fold CV criterion is formulated to be

$$CV_J(\mathbf{w}) = \sum_{j=1}^{J} \sum_{s=1}^{S} \left\{ \widetilde{F}^{(-j)}\left(x_{(j-1)S+s}, \mathbf{w}\right) - \bar{F}_{(j)}\left(x_{(j-1)S+s}\right) \right\}^2.$$

In other words, we consider the squared difference between the model-average estimator constructed *without* the jth group of real-world data and the ED constructed from *only* the jth group, summed across all groups. The empirically optimal weight vector resulting from this criterion is

$$\widehat{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{W}} CV_J(\mathbf{w}),$$

leading to the model-average estimator $\widehat{F}(x, \widehat{\mathbf{w}})$ of $F^c(x)$.

The optimization problem we need to solve to find $\widehat{\mathbf{w}}$ can be formulated as a quadratic program (QP); see Jiang and Nelson (2018) for the formulation and Nocedal and Wright (2006) for solving QPs. Specifically,

$$\begin{aligned}
\text{minimize}: \; & CV_J(\mathbf{w}) \\
&= \sum_{j=1}^{J} \sum_{s=1}^{S} \left\{ \widetilde{F}^{(-j)}\left(x_{(j-1)S+s}, \mathbf{w}\right) - \bar{F}_{(j)}\left(x_{(j-1)S+s}\right) \right\}^2 \\
&= \sum_{j=1}^{J} \sum_{s=1}^{S} \left\{ \sum_{m=1}^{q} w_m \left(\widetilde{F}_m^{(-j)}\left(x_{(j-1)S+s}\right) - \bar{F}_{(j)}\left(x_{(j-1)S+s}\right) \right) \right\}^2 \\
&= \sum_{j=1}^{J} \sum_{s=1}^{S} \left\{ \sum_{m=1}^{q} w_m c_{mjs} \right\}^2 \\
&= \sum_{j=1}^{J} \sum_{s=1}^{S} \mathbf{w}^\top \mathbf{C}_{js} \, \mathbf{w} = \mathbf{w}^\top \mathbf{C} \, \mathbf{w} \\
\text{subject to}: \; & \sum_{m=1}^{q} w_m = 1 \\
& w_m \ge 0, \; m = 1, 2, \ldots, q,
\end{aligned}$$

where the matrices $\mathbf{C}_{js}$ and $\mathbf{C}$ are defined in the obvious way. If the $q \times q$ matrix $\mathbf{C}$ is positive definite, then the objective function is strictly convex and the QP has a unique optimal solution; we refer to this as the *empirically optimal* model average. That each matrix $\mathbf{C}_{js}$ is positive semidefinite is clear from their construction. When X is continuous-valued and at least one of the candidate distributions is continuous, we show in Online Appendix C that the probability of there existing a $\mathbf{w}' \neq \mathbf{0}$ for which $(\mathbf{w}')^\top \mathbf{C} \, \mathbf{w}' = 0$ is zero; therefore, $\mathbf{C}$ is positive definite almost surely. The QP is easily solved via standard methods (Nocedal and Wright 2006). Notice that the construction of $\mathbf{C}$ is a one-time calculation and that QPs of such small size ($q \le 40$) can be solved very efficiently. The fitting algorithm is implemented in our R package FMAdist (https://cran.r-project.org/package=FMAdist).

Remark 3. Model averaging is intuitively appealing, as it enlarges the space of input model choices beyond $\mathcal{F}$ while still including the individual candidate distributions in $\mathcal{F}$ as special cases; it employs cross-validation as a robust method for fitting; and the empirically optimal solution is easy to find under weak assumptions. Under more restrictive assumptions, we show in Section 3.3 that this empirically optimal choice is in fact the best possible choice as $N \to \infty$.

3.2. Relationships to Other Input Modeling Methods

The greatest progress in input uncertainty quantification to date has been when the distribution family of F^c is assumed known [e.g., gamma(α, β)], so that the only uncertainty comes from the parameter estimates (e.g., $\widehat{\alpha}, \widehat{\beta}$). In our opinion, it is not universally the case

that real-world data both conform perfectly to a parametric distribution and are measured to sufficient precision to be indistinguishable, even though parametric distributions are often good approximations.

There has been some work on input uncertainty quantification that allows for distribution family uncertainty, specifically Chick (2001) and Zouaoui and Wilson (2004). Both papers take a Bayesian perspective, placing a prior distribution on the correct model family (e.g., exponential, Weibull, gamma), as well as each distributions' parameters, and derive the posterior distributions given $\mathbf{x}_N$. Although variate generation of inputs is identical to our method—first choosing the distribution family from the posterior, then generating variates—their goal is to fully represent input uncertainty in the posterior predictive distribution of the output Y, rather than trying to reduce it as we do; in fact, we provide no estimate of input uncertainty.

Another appealing solution is to use a parametric function $\widehat{F}$ that has the flexibility to get close to any F^c, and many distributions have been created for this purpose, including the generalized lambda distribution (Karian and Dudewicz 2000) to match moments or percentiles, as well as the Bézier distribution (Wagner and Wilson 1996), which can have an arbitrary number of parameters. However, these distributions were created to be flexible rather than to conform to particular process physics, leading to the possibility of overfitting or manifesting unusual features that are not consistent with the data. There is, after all, a reason that the standard arsenal of normal, lognormal, logistic, Weibull, gamma, Pareto, and so on continue to be used: Their existence is implied by theory that can hold approximately in practice.

As described in the previous section, the input model averaging approach allows us to exploit these tried-and-true families and also extend their reach through averaging. To resist overfitting, we use CV to select the weights; CV ensures that the weights do not give an average that is inconsistent with the distribution of the data, and the empirically optimal weights are unique and easily found.

3.3. Asymptotic Properties of Input Model Averages

In this section, we establish asymptotic properties of the empirically optimal model average under certain restrictions on the true distribution F^c and the individual candidate distributions in $\mathcal{F}$, and whether the ED $\bar{F}$ is in $\mathcal{F}$. As $N \to \infty$, (a) for certain classes of candidate distributions $\mathcal{F}$, when neither F^c nor $\bar{F}$ are in $\mathcal{F}$ individually, the empirically optimal model-average weights become the squared-error-optimal weights; and (b) when $\bar{F}$ is included in $\mathcal{F}$, its weight converges to 1. The first result implies that, under certain conditions, cross-validation provides the best-possible weights when no candidate distribution fits perfectly, whereas the second implies that it is consistent for F^c if we include the ED as a candidate.

We first establish the restrictions. Let $\boldsymbol{\beta}_m$ be the unknown parameter vector in the mth candidate distribution, and let $\widehat{\boldsymbol{\beta}}_m$ be its MLE for $m = 1, 2, \ldots, q$, which we assume exists. It is worth noting that $\widehat{\boldsymbol{\beta}}_m$ $(1 \le m \le q)$ is determined from each candidate distribution individually and not by the optimized linear combination of distributions. Further, let $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_1^{\mathrm{T}}, \ldots, \widehat{\boldsymbol{\beta}}_q^{\mathrm{T}})^{\mathrm{T}}$ with dimension κ. We require that the size q of the candidate set is finite. Furthermore, we assume that the following conditions hold:

(i) For each $x \in \mathbb{R}$, the density function $f_m(x; \boldsymbol{\beta}_m)$ of the *mth* candidate distribution is continuous at every $\boldsymbol{\beta}_m$ in the corresponding compact parameter space Θ_m.

(ii) There exists $\mathrm{E}[\log f^c(x)]$ and $|\log f_m(x; \boldsymbol{\beta}_m)| < l(x)$, where $l(x)$ is integrable with respect to F^c.

(iii) There exists a vector $\boldsymbol{\beta}_m^*$ at which the Kullback–Leibler information $\int_{\mathbb{R}} \log[f^c(x)/f_m(x; \boldsymbol{\beta}_m)] | f^c(x) dx$ attains a unique minimum.

Under these conditions, $\widehat{\boldsymbol{\beta}} \to \boldsymbol{\beta}^* = (\boldsymbol{\beta}_1^{*\mathrm{T}}, \ldots, \boldsymbol{\beta}_q^{*\mathrm{T}})^{\mathrm{T}}$ almost surely as $N \to \infty$; that is, the MLEs converge even when the distributions are misspecified. White (1982) further showed that

$$\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* = O_p(N^{-1/2}). \tag{6}$$

We assume that (6) is in force. The validity of (6) depends on conditions (i)–(iii), as well as assumptions A4, A5, and A6 of White (1982).

Remark 4. The canonical parameter space for many standard distributions is not compact, as assumed in (i), for example, for the normal distribution $\sigma > 0$. However, as a practical matter, assuming that there exists a large but compact space in which each parameter lies, for example, $\sigma \in [10^{-10}, 10^{10}]$, is reasonable since there is no requirement that the bound be known. In this sense, all of the distributions that have a density in the examples in Section 4 satisfy this condition.

We next define the notations needed to state our main results. Let $\mathbf{F}_0 = (F^c(x_1), \ldots, F^c(x_N))^{\mathrm{T}}$, the values of the true cdf evaluated at the data points, and let $\widehat{\mathbf{F}}_m = (\widehat{F}_m(x_1), \ldots, \widehat{F}_m(x_N))^{\mathrm{T}}$, the corresponding quantity for the *mth* candidate fitted distribution, for $m = 1, 2, \ldots, q$. We assume that the ED is not one of the q candidates. For any fixed $\mathbf{w}$, define a corresponding vector of values for the averaged distribution, with parameters fitted from data $\widehat{\mathbf{F}}(\mathbf{w}) = (\widehat{F}(x_1, \mathbf{w}), \ldots, \widehat{F}(x_N, \mathbf{w}))^{\mathrm{T}}$ and with the limiting parameters $\mathbf{F}^*(\mathbf{w}) = \widehat{\mathbf{F}}(\mathbf{w}) \,|_{\hat{\beta}=\beta^*}$.

Recall that CV leaves out sets of S consecutive data values in turn. In $\widetilde{\mathbf{F}}(\mathbf{w}) = (\widetilde{F}^{(-1)}(x_1, \mathbf{w}), \ldots, \widetilde{F}^{(-1)}(x_S, \mathbf{w}), \widetilde{F}^{(-2)}(x_{S+1}, \mathbf{w}), \ldots, \widetilde{F}^{(-J)}(x_N, \mathbf{w}))^{\mathrm{T}}$, we collect the cdf values for each data point based on the model average that

excludes it; that is, $\bar{\mathbf{F}} = (\bar{F}_{(1)}(x_1), \ldots, \bar{F}_{(1)}(x_S), \bar{F}_{(2)}(x_{S+1}), \ldots, \bar{F}_{(J)}(x_N))^{\mathrm{T}}$ is the corresponding vector using the ED. We assume that J is fixed, so that $S \to \infty$ as $N \to \infty$.

Finally, define the discrepancy $L_N^*(\mathbf{w}) = \|\mathbf{F}^*(\mathbf{w}) - \mathbf{F}_0\|^2$, and let $\xi_N = \inf_{\mathbf{w}\in\mathcal{W}} L_N^*(\mathbf{w})$ (with all weights assigned to distributions other than the ED).

For proving the results, we need the following regularity conditions.

Condition 1. *There exists a neighborhood $\mathcal{N}$ of $\boldsymbol{\beta}^*$ such that*

$$\sup_{\tilde{\boldsymbol{\beta}}\in\mathcal{N}} \left\| \frac{\partial \widehat{F}(x_i, \mathbf{w})}{\partial \widehat{\boldsymbol{\beta}}}\bigg|_{\hat{\boldsymbol{\beta}}=\tilde{\boldsymbol{\beta}}} \right\| = O_p(1)$$

uniformly for $i = 1, 2, \ldots, N$ and $\mathbf{w} \in \mathcal{W}$.

Condition 2. *For all $\mathbf{w} \in \mathcal{W}$, $N^{-1/2}\|\widehat{\mathbf{F}}(\mathbf{w}) - \widetilde{\mathbf{F}}(\mathbf{w})\|^2 = O_p(1)$, and $N^{-1/2}\{\widehat{\mathbf{F}}(\mathbf{w}) - \widetilde{\mathbf{F}}(\mathbf{w})\}^T\{\widehat{\mathbf{F}}(\mathbf{w}) - \bar{\mathbf{F}}\} = O_p(1)$.*

Condition 3. *When $N \to \infty$, there exists a sequence $c_N \to 0$ such that $\xi_N^2 \geq N/c_N$ almost surely.*

Condition 3 is well defined even if F^c is a nontrivial mixture of two or more elements of the candidate set $\mathcal{F}$. It can be seen that

$$\begin{aligned}\widehat{\mathbf{F}}(\mathbf{w}) - \widetilde{\mathbf{F}}(\mathbf{w}) = \Bigg(& \sum_{m=1}^{q} w_m \left\{ \widehat{F}_m(x_1) - \widetilde{F}_m^{(-1)}(x_1) \right\}, \ldots, \\ & \sum_{m=1}^{q} w_m \left\{ \widehat{F}_m(x_S) - \widetilde{F}_m^{(-1)}(x_S) \right\}, \\ & \sum_{m=1}^{q} w_m \left\{ \widehat{F}_m(x_{S+1}) - \widetilde{F}_m^{(-2)}(x_{S+1}) \right\}, \ldots, \\ & \sum_{m=1}^{q} w_m \left\{ \widehat{F}_m(x_N) - \widetilde{F}_m^{(-J)}(x_N) \right\} \Bigg)^T. \end{aligned} \tag{7}$$

Hence, Condition 2 requires the difference between the regular and leave-S out estimators to decrease sufficiently quickly as N increases. On the other hand, Condition 3 requires that ξ_N grows at a rate no slower than $N^{1/2}$. This in turn implies that the correct input distribution F^c must not be among the candidate distributions in the model average.

Theorem 1. *If $F^c \notin \mathcal{F}$, $\bar{F} \notin \mathcal{F}$, and Conditions 1–3 hold, then as the real-world sample size $N \to \infty$,*

$$\frac{\sum_{i=1}^{N} \left[\widehat{F}(x_i, \widehat{\mathbf{w}}) - F^c(x_i) \right]^2}{\inf_{\mathbf{w}\in\mathcal{W}} \sum_{i=1}^{N} \left[\widehat{F}(x_i, \mathbf{w}) - F^c(x_i) \right]^2} \xrightarrow{P} 1. \tag{8}$$

The proof is in Online Appendix A. Notice that the numerator and denominator are the sum of squared deviations of the model-average estimator from the true distribution, a comprehensive measure of fit. In the numerator, the weight $\widehat{\mathbf{w}}$ is obtained via J-fold CV with the empirical distribution, whereas in the denominator the minimum possible squared deviation weight is chosen. The result shows that, as the sample size increases, J-fold CV yields error *no larger than the minimum possible error* with the given set of candidate distributions, which we would expect to be smaller than choosing any *one* distribution from $\mathcal{F}$ when $F^c \notin \mathcal{F}$. Notice that the condition "$F^c \notin \mathcal{F}$" does not prohibit F^c from being a nontrivial mixture of two or more elements of $\mathcal{F}$.

Theorem 1 does not establish that $\widehat{F}(x_i, \widehat{\mathbf{w}})$ is asymptotically consistent for F^c. However, as noted earlier, the ED is unbiased and consistent for F^c. Therefore, we also consider including $\bar{F}$ in the candidate set $\mathcal{F}$ for model averaging. Although (8) no longer holds, model averaging becomes consistent in the sense that, in the limit, all of the weight is on the ED.

Theorem 2. *If $F^c \notin \mathcal{F}$ but $\bar{F} \in \mathcal{F}$, and if Conditions 1–3 hold, then $\widehat{w}_{ED} \xrightarrow{P} 1$ as $N \to \infty$.*

The proof is in Online Appendix B. The effect of including $\bar{F}$ is that the other parametric distributions smooth the ED and provide better tail behavior. This is important because the ED being unbiased does not imply that the output $\bar{Y} \sim F_Y(y \mid \bar{F})$ has the same distribution, or even the same mean, as the ideal output Y^c because the simulation is in general a highly nonlinear transformation of inputs to outputs. See the commentary in Song et al. (2015). Nevertheless, we will show empirically that the ED of the entire data set, $\bar{F}$, is often a good choice when the criterion is recovering the distribution of Y^c, and model averaging with the ED can be superior to either the ED alone or model averaging of parametric distributions.

Remark 5. Among the list of possible candidates $\mathcal{F}$ could be kernel density estimators and the more-recent log-concave density estimators (Cule et al. 2010, Kim and Samworth 2016). These semiparametric methods do not leverage process physics—an advantage of our approach—but do have excellent convergence rates to the true distribution. However, we would expect our fits to be smoother for small to moderate N. That said, these methods are not natural candidates for our model averaging because they directly estimate the density, whereas we require the cdf.

4. Experiments

In this section, we evaluate our proposal empirically. Recall that our interest is in how properties of the simulation output $\widehat{Y} \sim F_Y(y \mid \widehat{F})$ compare with the ideal output $Y^c \sim F_Y(y \mid F^c)$ (Section 4.1) and also how our fitted model-average distribution $\widehat{F}(x; \widehat{\mathbf{w}})$ compares to the distribution that generated the input data F^c (Section 4.2). The assessments in this section are

quantitative; see Jiang and Nelson (2018) for some graphical illustrations of the attained fits and Online Appendix D for additional documentation.

We reach the following broad conclusions: Model averaging, especially including the ED as a candidate, is often substantially superior to any *single* choice from $\mathcal{F}$, and typically no worse. Given a large number q of candidate distributions, it is best to screen out obviously poor choices first and do model averaging over a smaller subset of $\mathcal{F}$. When either the size of the real-world data sample N is large, or none of the candidate distributions has the capability to fit well (e.g., data are multimodal but choices in $\mathcal{F}$ are all unimodal), the weight on the ED, $\widehat{w}_{ED}$, tends to be large. Thus, the ED provides protection against a badly chosen candidate set (which can occur when using the built-in set in a software package) and provides the consistency established by Theorem 2. Although not specifically targeted in our experiments, it seems clear that employing a candidate set with common, appropriate support, and containing candidates that are well justified by process physics when available, is helpful. Finally, we found no systematic difference from using $J = 5$ or 10 folds for CV; we would never use more than 10 folds and recommend $J = 5$ when N is small.

4.1. Evaluation of the Output Distribution

To evaluate the various methods with respect to the output distribution, we use the relative distribution method of Handcock and Morris (2006). A brief explanation of the method follows: Let the distributions of $\widehat{Y}$ and Y^c be denoted by $F_{\widehat{Y}}(y)$ and $F_{Y^c}(y)$, respectively. Define the grade transformation of $\widehat{Y}$ to Y^c as

$$U = F_{Y^c}\left(\widehat{Y}\right), \tag{9}$$

obtained by transforming $\widehat{Y}$ by F_{Y^c}. The CDF of U can be expressed as

$$G(u) = F_{\widehat{Y}}\left(F^{-1}_{Y^c}(u)\right) \tag{10}$$

for $0 \le u \le 1$, where $F^{-1}_{Y^c}(u) = \inf\{y \mid F_{Y^c}(y) \ge u\}$ is the quantile function of F_{Y^c}.

It is easily seen that if $\widehat{Y} \overset{\mathcal{D}}{=} Y^c$, then the CDF of U is a 45° line. When $\widehat{Y} \overset{\mathcal{D}}{\neq} Y^c$, then the closer $G(u)$ is to the 45° line, the better the fit provided by $\widehat{Y}$. In our analysis, we use the unsigned area between $G(u)$ and the 45° line over $0 \le u \le 1$ to evaluate the effectiveness of the method. We denote the gap by $A(u) = |G(u) - u|$, so that the area is $A = \int_0^1 A(u)\,du$. Clearly, we could compare a list of individual properties, such as the mean and variance, but A provides a comprehensive measure of performance. The relative distribution method is in the same spirit as the tail-probability plot method (see, e.g., Heyde and Kou 2004). When the true distribution of Y^c is not available, as in most of our examples, it is represented by a very large sample from F_{Y^c}; this is possible for us because the distributions of the inputs F^c are known. To implement the relative distribution analysis, we used the codes at https://csde.washington.edu/~handcock/RelDist/Software/R/.

The following are features of our empirical evaluation:

- We apply input model averaging to cases of the SAN, $GI/G/1$ and HRS simulations, as described in Sections 2.1–2.3, for different quantities of real-world data, N, used to fit the input models. We generate the "real-world data" from fully specified distributions.
- We consider instances in which the candidate set $\mathcal{F}$ does, or does not, contain the true input distributions F^c. The single "best-fit" distribution, which represents common input-modeling practice, is selected from this set both by minimum AIC and minimum BIC.
- We refer to our frequentist model averaging method as *JFCV* (for J-fold cross-validation). The ED is considered both as an individual input distribution method and a candidate within the JFCV model average. We refer to the JFCV method that includes the ED as a default candidate as the JFCV(ED) method. Thus, our five competing methods are AIC, BIC, ED, JFCV, and JFCV(ED).
- When evaluating the performance of the methods, we consider both the area A discussed above and the tail area $A_{\text{tail}} = \int_{0.9}^1 A(u)\,du$. We are interested in A_{tail} because there is a common belief that the ED, which does not model the tail of the distribution beyond the largest data point in the real-world sample, may be an inferior method when interest centers on the tail of the simulation output Y. Each experiment is repeated for 100 macroreplications and the results reported are averages of A or A_{tail} across these 100 macroreplications. When presenting the results, we usually normalize the average area (of A or A_{tail} depending on the focus of interest) generated by the JFCV method to 1, although in some cases we also present the raw average area. Hence, if the relative average area produced by a method is larger than 1, then it is inferior to JFCV, and vice versa, based on this metric. All results are displayed to statistically meaningful digits of precision. We also examine how the JFCV weight $\widehat{\mathbf{w}}$ changes as N increases.
- For the SAN experiment, we also very precisely estimate the mean squared error (MSE) of a point estimate for the probability of late project completion.

4.1.1. SAN Experiment. We begin with the SAN described in Section 2.1, for which there are five input distributions for the five activity times, $X_1, X_2, \ldots, X_5$. For cases I–III, the true distributions are made up of mixture distributions that are *not* contained in any of the candidate sets, whereas cases IV–V include distributions that are contained in the candidate sets.

Results for cases I–V are reported in Online Appendix D. Here we report case VI, which uses a candidate set $\mathcal{F}$ that is common to all commercial distribution fitting products: $\mathcal{F}_4$ = {normal, lognormal, beta, exponential, gamma, Weibull} plus possibly the ED. In addition to the candidate set $\mathcal{F}_4$, we also consider a smaller subset within it containing the "best" three based on minimum AIC and BIC selections. We refer to this subset as $\mathcal{F}_4^{(3)}$ and apply the JFCV methods under this subset as well as the full set $\mathcal{F}_4$. In the event that AIC and BIC do not lead to the same set of best distributions, averaging under $\mathcal{F}_4^{(3)}$ will involve more than three distributions.

The true activity-time distributions are Pareto, Rayleigh, and loglogistic, as shown in Table 1. In each case, the mean activity time is approximately 1. None of these are contained in the candidate set.

The results are displayed in Table 2. JFCV is superior to any single choice made via best AIC or BIC, and JFCV(ED), which includes the ED in the candidate set, is substantially better than JFCV alone. In this example selecting a subset of the top three distributions before modeling averaging has little or no effect; however, in Online Appendix D, we show that it can be useful in other scenarios for the SAN, as well as in our distribution-to-distribution comparisons in Section 4.2.

Although A provides a comprehensive measure of output-distribution performance, we also display some results for the MSE of a point estimate of $\Pr\{Y^c > 6.65\}$, since the probability of completing beyond a due date is often important in project planning; 6.65 is the 0.9 quantile (based on a side experiment with one million replications). This case is to give a sense of the effect of modeling averaging on point-estimator performance. Table 3 displays results for real-world sample sizes $N = 100, 1{,}000$, and $R = 1{,}000$ replications of the SAN; a large number of replications are required so that point estimator variance does not overwhelm the bias reduction that we hope to be revealed. The MSE is estimated from 5,000 macroreplications of the entire experiment, and the standard error of the estimate is also displayed. We see that when N is small, model averaging yields substantial improvement over the ED or best AIC choices; when N is larger, the ED and model average are indistinguishable.

4.1.2. $GI/G/1$ Experiment. Next we examine results for two cases of the $GI/G/1$ queue described in Section 2.2: An $M/M/1$ queue (case VII), meaning exponential interarrival and service times, and a $GI/G/1$ queue with balanced hyperexponential interarrival times,

$$X_1 \sim \begin{cases} \text{exponential}(1) & \text{with probability } 1/2 \\ \text{exponential}(20) & \text{with probability } 1/2, \end{cases}$$

and service times that follow the mixture distribution,

$$X_2 \sim \begin{cases} \text{unif}(10, 20) & \text{with probability } 2/5 \\ \text{gamma}(2.875, 1/2) & \text{with probability } 3/5, \end{cases}$$

which we label as case VIII. In both cases, the implied traffic intensity is $\mathrm{E}(X_2)/\mathrm{E}(X_1) = 0.9$, and the output we consider is the waiting time of the fifth arrival Y_5. We consider candidate sets

$$\begin{aligned} \mathcal{F}_1 &= \{\text{truncated normal, beta, gamma}\}, \\ \mathcal{F}_2 &= \mathcal{F}_1 \cup \{\text{lognormal, Weibull}\}, \\ \mathcal{F}_3 &= \mathcal{F}_2 \cup \{\text{negative binomial, discrete uniform,} \\ &\qquad \text{Poisson, continuous uniform, loglogistic,} \\ &\qquad \text{inverse Gaussian, Pareto, binomial}\}. \end{aligned}$$

Tables 4 and 5 contain the $M/M/1$ results for relative average A and A_{tail}, respectively. For capturing the entire output distribution of Y_5 as measured by A, JFCV and JFCV(ED) tend to be better than AIC, BIC, and ED, even though the true exponential distribution is in all candidate sets $\mathcal{F}_1$–$\mathcal{F}_3$ in the form of the gamma distribution, and again in sets $\mathcal{F}_2$–$\mathcal{F}_3$ in the form of the Weibull distribution. AIC and BIC improve substantially in capturing the tail behavior as measured by A_{tail} but do not beat JFCV(ED).

Tables 6 and 7 present corresponding results for the $GI/G/1$. When interest centers on A, the JFCV(ED) method is the clear favorite, followed by the ED, which has an edge over the JFCV, which in turn delivers better performance than AIC and BIC in the majority of cases. When interest centers on A_{tail}, JFCV(ED) remains the best, the ED can yield worse

Table 1. True Activity-Time Distributions for Case VI SAN Example

Activity	Distribution	Parameters	CDF
X_1	Rayleigh	$\sigma = \pi/2$	$1 - \exp(x^2/(2\sigma^2)), x \geq 0$
X_2	Pareto	$\mu = 1/4, \sigma = 3/16, \xi = 3/4$	$1 - (1 + \xi(x-\mu)/\sigma)^{-1/\xi}, x \geq \mu$
X_3	Pareto	$\mu = 1/2, \sigma = 1/4, \xi = 1/2$	$1 - (1 + \xi(x-\mu)/\sigma)^{-1/\xi}, x \geq \mu$
X_4	Loglogistic	$\alpha = 0.23, \beta = 1.21$	$\dfrac{1}{1 + (x/\alpha)^{-\beta}}, x \geq 0$
X_5	Loglogistic	$\alpha = 2/\pi, \beta = 2$	$\dfrac{1}{1 + (x/\alpha)^{-\beta}}, x \geq 0$

Table 2. Numerical Results for SAN Experiment Case VI

	Actual average A					Relative average A				
Scenario	JFCV	AIC	BIC	ED	JFCV (ED)	JFCV	AIC	BIC	ED	JFCV (ED)
$\mathcal{F}_4$	0.05	0.06	0.06	0.06	0.04	1.00	1.07	1.07	1.06	0.77
$\mathcal{F}_4^{(3)}$	0.06	0.06	0.06	0.06	0.04	1.00	1.05	1.05	1.05	0.76

performance than the JFCV, and the AIC and BIC selections can be particularly bad when there is a large set of candidate distributions.

4.1.3. HRS Experiment. Finally, we consider the HRS example of Section 2.3. We consider the following setup, labeled as case IX in our subsequent presentation of results. Let the inputs be

$$X_1 \sim \begin{cases} \text{unif}(0,1) & \text{with probability } 0.5 \\ \text{exponential}(1) & \text{with probability } 0.5 \end{cases}$$
$$X_2 \sim \begin{cases} \text{N}(100,100) & \text{with probability } 0.5 \\ \text{gamma}(20,0.2) & \text{with probability } 0.5. \end{cases}$$

Notice that $\text{E}(X_1) = 1$ and $\text{E}(X_2) = 100$. Recall that Y is thought of as the time to failure.

Tables 8 and 9 present the results. This is a very difficult problem for which the distribution of Y is very sensitive to the input distributions. This makes the performance of JFCV(ED) impressive, as it is clearly the best across all cases. The performance of the JFCV, AIC, BIC, and ED methods is rather diverse. AIC's performance is either on a par with, or slightly better than, BIC. None of the JFCV, AIC, BIC and ED can strictly dominate the others, although JFCV tends to be the winner when considering A.

4.2. Evaluation of the Input Distribution

In this section, we present results that directly assess the quality of the model-average fit $\hat{F}(x;\widehat{\mathbf{w}})$ with respect to the true distribution F^c. In the unlikely event that $F^c \in \mathcal{F}$, one should not expect model averaging to do better since an empirical weight of precisely 1 assigned to any particular distribution, including F^c, is a probability 0 outcome. Therefore, we focus on cases in which $F^c \notin \mathcal{F}$.

Table 3. MSE Results for SAN Experiment, Case VI, for Estimating $\Pr\{Y^c > 6.65\}$

N	R	Candidates	MSE	SE (MSE)
100	1,000	ED	0.00254	4.8E-05
100	1,000	Best AIC	0.00116	1.6E-05
100	1,000	$\mathcal{F}_4$ + ED	0.00079	1.8E-05
1,000	1,000	ED	0.00016	2.9E-06
1,000	1,000	Best AIC	0.00093	8.2E-06
1,000	1,000	$\mathcal{F}_4$ + ED	0.00015	3.1E-06

Specifically, our candidate set is all or part of

$$\mathcal{F} = \{\text{normal, lognormal, exponential, Weibull, gamma, ED}\}$$

whereas F^c is Rayleigh, Pareto, generalized lambda (Karian and Dudewicz 2000), hyperexponential, or mixtures of these. For measures of fit, we compared the mean and variance of the fitted distributions to those of F^c (as a sanity check), but, more importantly, we computed the following:

Kolmogorov-Smirnov distance (K-S): $\max_x |\hat{F}(x;\widehat{\mathbf{w}}) - F^c(x)|$;

Cramér von-Mises distance (Cv-M): $\int [\hat{F}(x;\widehat{\mathbf{w}}) - F^c(x)]^2 \, dF^c(x)$;

Anderson-Darling distance (A-D): $\int \frac{[\hat{F}(x;\widehat{\mathbf{w}}) - F^c(x)]^2}{F^c(x)(1-F^c(x))} \, dF^c(x)$.

K-S examines the largest absolute gap between the cdfs; Cv-M and A-D are likelihood weighted squared areas between them, with A-D further emphasizing differences in the tails. We also recorded the weights assigned to each distribution in the model average. Real-world sample sizes of $N = 100, 1{,}000$ were employed, and all results were averaged over 1,000 macroreplications of the experiment.

We present results that represent the more-favorable and less-favorable performance of model averaging from this large number of cases. Not surprisingly, no approach dominates on all instances and all measures, so "favorable" is somewhat subjective. Overall, we found the following:

- Model averaging can improve over any single choice from $\mathcal{F}$, and the best model-average tends never to be worse.
- Including the ED in $\mathcal{F}$ is almost always valuable for measures other than K-S; ED alone often has the poorest K-S performance, which makes sense as it is a discrete approximation to a continuous F^c.
- Reducing the size of $\mathcal{F}$ to the top AIC/BIC choices before model averaging improves fit; often

Table 4. Numerical Results for $M/M/1$ Queue with $N = 100$

		Relative average A				
Case	Scenario	JFCV	AIC	BIC	ED	JFCV (ED)
VII	$\mathcal{F}_1$	1.00	1.04	1.04	1.00	0.98
	$\mathcal{F}_2$	1.00	1.04	1.04	1.01	0.98
	$\mathcal{F}_3$	1.00	1.00	0.99	0.96	0.94
	$\mathcal{F}_3^{(3)}$	1.00	1.09	1.08	1.05	1.00
	$\mathcal{F}_3^{(6)}$	1.00	1.07	1.06	1.02	0.98

Table 5. Numerical Results for $M/M/1$ Queue, Tail Estimation, with $N = 100$

		Relative average A_{tail}				
Case	Scenario	JFCV	AIC	BIC	ED	JFCV (ED)
VII	$\mathcal{F}_1$	1.00	0.74	0.74	1.24	0.84
	$\mathcal{F}_2$	1.00	0.71	0.71	1.17	0.71
	$\mathcal{F}_3$	1.00	0.43	0.42	0.65	0.47
	$\mathcal{F}_3^{(3)}$	1.00	0.82	0.81	1.26	0.77
	$\mathcal{F}_3^{(6)}$	1.00	0.53	0.53	0.82	0.51

model averaging the single best fit and the ED is the consensus best choice.

- The more distinct F^c is from any other choice in $\mathcal{F}$, the more weight is applied to the ED; for instance, this occurred when we created a bimodel true distribution F^c via a mixture (all of the candidates in $\mathcal{F}$ are unimodal).
- In a targeted test to study the effect of nested distributions, we found that using $\mathcal{F} = \{\text{exponential, Weibull, gamma}\}$ for model averaging when F^c is exponential leads to a noticeably poorer fit than choosing any one of the candidates. A tentative recommendation is to avoid nesting, such as including exponential and Erlang in a set that already includes Weibull and gamma.

4.2.1. More-Favorable Performance. Here F^c is Rayleigh with parameter 0.5, from which we have $N = 1{,}000$ observations, with full candidate set $\mathcal{F} = \{\text{normal, lognormal, exponential, gamma, ED}\}$, and we use $J = 5$ folds for fitting the weights. The gamma distribution is always the best AIC fit. Results are shown in Table 10. Either gamma + ED or using all of $\mathcal{F}$ provide arguably the best fits based on our three performance measures. For the same experiment with only $N = 100$ "real-world" observations, gamma + ED was the best choice, and better than model averaging larger sets. This suggests that when the quantity of input data is small it is even more important to first screen the larger set $\mathcal{F}$ before model averaging.

For a second favorable example, F^c is Pareto with location parameter 1 and shape parameter 3, from which we have $N = 100$ observations, with full candidate set $\mathcal{F} = \{\text{normal, lognormal, gamma, Weibull, ED}\}$, and we use $J = 5$ folds for fitting the weights. Either the gamma, lognormal, or Weibull distribution was chosen as the best AIC fit in some macroreplication, so we included them all as individual choices. Results are shown in Table 11. Individually, the Weibull provides a good fit in this case, yet improvement is still possible by model averaging a smaller set of distributions than the full set.

Although not shown here because the result is obvious, model averaging including the ED had very favorable performance relative to any single choice when F^c was obtained by a mixture (e.g., of two Rayleigh's with different parameters) so as to create a bimodel distribution; in such cases, the ED received a weight of around 0.9. This illustrates that model averaging with the ED provides protection against a poorly chosen candidate set, which might occur if distribution fitting was automated. Of course, bimodal and mixture distributions can be included as candidates.

4.2.2. Less-Favorable Performance. In all of our experiments, there was *some* model-average distribution that did as well or better than any single choice, but, in a few cases, this was very sensitive to the distributions chosen as candidates; the most extreme case follows.

In this example, F^c is a generalized lambda distribution with $\lambda_1 = 3$, $\lambda_2 = 2$, $\lambda_3 = 1.5$, and $\lambda_4 = 0.5$. With these choices, the density has a bathtub shape. On each of 1,000 macroreplications, we obtained $N = 100$ observations, with full candidate set $\mathcal{F} = \{\text{normal, lognormal, exponential, gamma, Weibull, ED}\}$, and we

Table 6. Numerical Results for $GI/G/1$ Queue with $N = 100$

		Relative average A				
Case	Scenario	JFCV	AIC	BIC	ED	JFCV (ED)
VIII	$\mathcal{F}_1$	1.00	0.94	0.94	0.90	0.83
	$\mathcal{F}_2$	1.00	1.05	1.05	0.96	0.88
	$\mathcal{F}_3$	1.00	1.15	1.16	0.81	0.78
	$\mathcal{F}_3^{(3)}$	1.00	1.30	1.31	0.92	0.84
	$\mathcal{F}_3^{(6)}$	1.00	1.24	1.25	0.88	0.82

Table 7. Numerical Results for $GI/G/1$ Queue, Tail Estimation, with $N = 100$

		Relative average A_{tail}				
Case	Scenario	JFCV	AIC	BIC	ED	JFCV (ED)
VIII	$\mathcal{F}_1$	1.00	0.64	0.64	0.99	0.58
	$\mathcal{F}_2$	1.00	0.82	0.82	1.35	0.67
	$\mathcal{F}_3$	1.00	1.40	1.42	0.68	0.37
	$\mathcal{F}_3^{(3)}$	1.00	2.49	2.54	1.22	0.59
	$\mathcal{F}_3^{(6)}$	1.00	1.66	1.69	0.81	0.38

Table 8. Numerical Results for HRS with $N = 100$

		Relative average A				
Case	Scenario	JFCV	AIC	BIC	ED	JFCV (ED)
IX	$\mathcal{F}_1$	1.00	1.00	1.00	0.58	0.37
	$\mathcal{F}_2$	1.00	0.73	0.73	0.64	0.41
	$\mathcal{F}_3$	1.00	1.26	1.27	1.27	0.75
	$\mathcal{F}_3^{(3)}$	1.00	1.17	1.18	1.18	0.70
	$\mathcal{F}_3^{(6)}$	1.00	1.29	1.31	1.30	0.79

Table 9. Numerical Results for HRS, Tail Estimation, with $N = 100$

		Relative average A_{tail}				
Case	Scenario	JFCV	AIC	BIC	ED	JFCV (ED)
IX	$\mathcal{F}_1$	1.00	1.00	1.00	0.66	0.55
	$\mathcal{F}_2$	1.00	0.85	0.85	0.70	0.57
	$\mathcal{F}_3$	1.00	0.74	0.75	0.70	0.51
	$\mathcal{F}_3^{(3)}$	1.00	1.20	1.22	1.13	0.82
	$\mathcal{F}_3^{(6)}$	1.00	0.83	0.85	0.79	0.61

used $J = 5$ folds for fitting the weights. The lognormal was chosen as the best AIC fit, but we explored other combinations as well. Results are shown in Table 12. Notice that averages of lognormal + ED and lognormal + gamma + normal + ED offer significant improvement on all measures over the single lognormal choice, but lognormal + gamma + normal and the full set $\mathcal{F}$ have inferior A-D statistics.

5. Conclusions

Model risk due to input uncertainty arises because the fitted input distribution $\widehat{F}$ deviates from the true distribution of the input data F^c. When F^c is known to belong to a certain parametric family, it makes sense to use statistically efficient parameter estimates, which would often be the MLEs. Many methods for quantifying the impact of input parameter uncertainty on simulation performance estimates for this case have been proposed.

However, at best we should expect a standard parametric family to be a good approximation for F^c, which means that there is error that does not disappear, even as the real-world input sample size $N \to \infty$. When the input data are also used to select the family, as is common practice, the possible error is compounded.

In this paper, we proposed using frequentist model averaging as an innovative way to construct better input models, meaning input models that yield more faithful output performance. Since the optimal weights are unknown, we estimated them using J-fold cross-validation. We showed that under mild conditions the empirically optimal model average is unique and easily obtained and that under more restrictive conditions the empirically optimal weights yield the best possible weighted average distribution as the sample size increases. This method augments current input modeling practice and requires no alternation of the simulation model or additional simulation runs.

We also observed that the empirical distribution (ED) is frequently a very good input-modeling choice when the objective is to get close to the ideal output distribution, F_{Y^c}; this seems not to be very well known. Including the ED in the candidate set $\mathcal{F}$ for model averaging hedges against possible inadequacy of the ED, as occurred in some of our examples, especially when tail behavior of Y is of interest. The JFCV(ED) input models were often the best by a significant margin, were always very good performers in our experiments, and seem to be a powerful addition to the standard input modeling pallet.

Table 10. Results from 1,000 Macroreplications for $N = 1{,}000$ Observations from a Rayleigh Distribution F^c

Distributions	w	K-S	Cv-M	A-D
Gamma	1	0.034	0.501	3.137
ED	1	0.028	0.172	1.021
Gamma + ED	(0.257, 0.743)	0.026	0.182	1.080
$\mathcal{F}$	(0.387, 0.052, 0.017, 0.479, 0.065)	0.018	0.148	1.941

Table 11. Results from 1,000 Macroreplications for $N = 100$ Observations from a Pareto Distribution F^c

Distributions	w	K-S	Cv-M	A-D
Lognormal	1	0.065	0.203	1.256
Gamma	1	0.064	0.206	1.259
Weibull	1	0.054	0.144	0.931
ED	1	0.085	0.164	0.990
Logn + ED	(0.748, 0.252)	0.064	0.164	1.002
Weibull + ED	(0.678 0.322)	0.060	0.137	0.849
Gamma + ED	(0.407, 0.593)	0.071	0.152	0.914
Logn + Weibull + ED	(0.571, 0.312, 0.117)	0.056	0.143	0.875
Logn + gamma + Weibull + ED	(0.575, 0.217, 0.098, 0.110)	0.056	0.142	0.872
$\mathcal{F}$	(0.073, 0.668, 0.039, 0.118, 0.103)	0.060	0.154	1.261

Table 12. Results from 1,000 Macroreplications for $N = 100$ Observations from a Generalized Lambda Distribution F^c

Distributions	w	K-S	Cv-M	A-D
Logn	1	0.090	0.266	4.119
Logn + ED	(0.170, 0.830)	0.080	0.162	1.079
Logn + gamma + normal	(0.810, 0, 0.190)	0.096	0.269	4.164
Logn + gamma + normal + ED	(0.091, 0, 0.011, 0.8)	0.079	0.163	1.091
$\mathcal{F}$	(0.014, 0.381, 0.130, 0.008, 0.008, 0.459)	0.111	0.231	9.118

Acknowledgments

The authors thank Eunhye Song for insights on averaging with the empirical distribution, and the area editor, associate editor, and referees for enlightening reviews.

Endnote

[1] Many new input models have been invented, particularly for multivariate and nonstationary inputs; the lack of progress to which we refer is in the methods for fitting these models to data.

References

Ankenman BE, Nelson BL (2012) A quick assessment of input uncertainty. Laroque C, Himmelspach J, Pasupathy R, Rose O, Uhrmacher AM, eds. *Proc. 2012 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 241–250.

Banks J, Carson J, Nelson B, Nicol D (2010) *Discrete-Event System Simulation* (Prentice Hall, Upper Saddle River, NJ).

Barton RR (2012) Tutorial: Input uncertainty in output analysis. Laroque C, Himmelspach J, Pasupathy R, Rose O, Uhrmacher AM, eds. *Proc. 2012 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 1–12.

Barton RR, Nelson BL, Xie W (2013) Quantifying input uncertainty via simulation confidence intervals. *INFORMS J. Comput.* 26(1): 74–87.

Burt JM, Garman MB (1971) Conditional Monte Carlo: A simulation technique for stochastic network analysis. *Management Sci.* 18(3): 207–217.

Cheng RCH, Holland W (1997) Sensitivity of computer simulation experiments to errors in input data. *J. Statist. Comput. Simulation* 57(1–4):219–241.

Cheng RCH, Holland W (1998) Two-point methods for assessing variability in simulation output. *J. Statist. Comput. Simulation* 60(3):183–205.

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

Corlu C, Biller B (2013) A subset selection procedure under input parameter uncertainty. Pasupathy R, Kim SH, Tolk A, Hill R, Kuhl ME, eds. *Proc. 2013 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 463–473.

Cule M, Samworth R, Stewart M (2010) Maximum likelihood estimation of a multi-dimensional log-concave density. *J. Royal Statist. Soc. Ser. B: Statist. Methodology* 72(5):545–607.

Fan W, Hong LJ, Zhang X (2013) Robust selection of the best. Pasupathy R, Kim SH, Tolk A, Hill R, Kuhl ME, eds. *Proc. 2013 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 868–876.

Ghosh S, Lam H (2015) Mirror descent stochastic approximation for computing worst-case stochastic input models. Yilmaz L, Chan WKV, Moon I, Roeder TMK, Macal C, Rossetti MD, eds. *Proc. 2015 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 425–436.

Glynn PW, Lam H (2018) Constructing simulation output intervals under input uncertainty via data sectioning. Rabe M, Juan A, Mustafee N, Skoogh A, Jain S, Johansson B, eds. *Proc. 2018 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 1551–1562.

Gross D, Shortle JF, Thompson JM, Harris CM (2008) *Fundamentals of Queueing Theory*, 4th ed. (Wiley, New York).

Handcock MS, Morris M (2006) *Relative Distribution Methods in the Social Sciences* (Springer, New York).

Hansen BE (2007) Least squares model averaging. *Econometrica* 75(4):1175–1189.

Hansen BE, Racine J (2012) Jackknife model averaging. *J. Econometrics* 167(1):38–46.

Heyde C, Kou S (2004) On the controversy over tailweight distributions. *Oper. Res. Lett.* 32(5):399–408.

Hjort NL, Claeskens G (2003) Frequentist model average estimators. *J. Amer. Statist. Assoc.* 98(464):879–899.

Jiang WX, Nelson BL (2018) Better input modeling via model averaging. Rabe M, Juan A, Mustafee N, Skoogh A, Jain S, Johansson B, eds. *Proc. 2018 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 1575–1586.

Karian ZA, Dudewicz EJ (2000) *Fitting Statistical Distributions: The Generalized Lambda Distribution and Generalized Bootstrap Methods* (CRC Press, New York).

Kim AK, Samworth RJ (2016) Global rates of convergence in log-concave density estimation. *Ann. Statist.* 44(6):2756–2779.

Lam H (2016) Input uncertainty and robust analysis in stochastic simulation: Advanced tutorial. Roeder TMK, Frazier PI, Szechtman R, Zhou E, Huschka T, Chick SE, eds. *Proc. 2016 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 179–192.

Lam H, Qian H (2018) Subsampling variance for input uncertainty quantification. Rabe M, Juan A, Mustafee N, Skoogh A, Jain S, Johansson B, eds. *Proc. 2018 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 1611–1622.

Law AM, Kelton WD (1991) *Simulation Modeling and Analysis*, 2nd ed. (McGraw-Hill, New York).

Liang H, Zou G, Wan ATK, Zhang X (2011) Optimal weight choice for frequentist model average estimators. *J. Amer. Statist. Assoc.* 106(495):1053–1066.

McLachlan G, Peel D (2004) *Finite Mixture Models* (John Wiley & Sons, New York).

Nocedal J, Wright SJ (2006) *Numerical Optimization*, 2nd ed. (Springer, New York).

Song E, Nelson BL (2015) Quickly assessing contributions to input uncertainty. *IIE Trans.* 47(9):1–17.

Song E, Nelson BL, Hong LJ (2015) Input uncertainty and indifference-zone ranking & selection. Yilmaz Chan, L WKV, Moon I, Roeder TMK, C Macal, Rossetti MD, eds. *Proc. 2015 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 414–424.

Song E, Nelson BL, Pegden CD (2014) Advanced tutorial: Input uncertainty quantification. Tolk A, Diallo S, Ryzhov I, Yilmaz L, Buckley S, Miller J, eds. *Proc. 2014 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 162–176.

Wagner M, Wilson J (1996) Using univariate Bézier distributions to model simulation input processes. *IIE Trans.* 28(9):699–711.

Wan ATK, Zhang X, Zou G (2010) Least squares model averaging by Mallows criterion. *J. Econometrics* 156(2):277–283.

White H (1982) Maximum likelihood estimation of misspecified models. *Econometrica* 50(1):1–25.

Whitt W (1981) Approximating a point process by a renewal process: The view through a queue, an indirect approach. *Management Sci.* 27(6):619–636.

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

Zhang X, Wan ATK, Zou G (2013) Model averaging by jackknife criterion in models with dependent data. *J. Econometrics* 174(2): 82–94.

Zhou E, Xie W (2015) Simulation optimization when facing input uncertainty. Yilmaz L, Chan WKV, Moon I, Roeder TMK, Macal C, Rossetti MD, eds. *Proc. 2015 Winter Simulation Conf.* (IEEE, Piscataway, NJ), 3714–3724.

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