Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

INQUIRE: INPUT-AWARE QUANTIFICATION OF UNCERTAINTY FOR INTERPRETATION, RISK, AND EXPERIMENTATION

Yujing Lin[1], Jingtao Zhang[2], Mitchell Perry[3], Xiaoyu Lu[2], Yunan Liu[3], and Hoiyi Ng[3]

[1]Amazon Supply Chain Optimization Technology, Austin, USA
[2]Amazon Supply Chain Optimization Technology, Bellevue, USA
[3]Amazon Supply Chain Optimization Technology, New York City, USA

ABSTRACT

Stochastic discrete event simulation is a vital tool across industries. However, the high dimensionality and complexity of real-world systems make it challenging to develop simulations that accurately model and predict business metrics for users when faced with inaccurate input and model fidelity limitations. Addressing this challenge is critical for improving the effectiveness of industrial simulations. In this work, we focus on simulation output uncertainty, a crucial summary statistic for assessing business risks. We introduce a novel framework called INput-aware Quantification of Uncertainty for Interpretation, Risk, and Experimentation (INQUIRE). At the heart of INQUIRE, we develop a residual-based uncertainty prediction model driven by key input parameters. Then we incorporate a skewness-detection procedure for quantile estimation that provides risk assessment. To analyze how input parameters evolution influences simulation output uncertainty, we introduce a Shapley-value-based interpretation method. Additionally, our framework enables more efficient simulation-driven experimentation, enhancing strategic decision-making by providing deeper insights.

1 INTRODUCTION

1.1 Background and Motivation

A stochastic *discrete event simulation* (DES) models stochastic behavior of a system by specifying probability distributions for random variables used in the model, and it is widely used to estimate counterfactual impact or to generate predictions for future events in industry. Business owners often create digital twins of their business systems through simulation, which models the dynamic interaction across different system components and relationship with macroeconomic evolution. DES is critical for business owners as it enables event-driven forward-looking forecasts for business management and "what-if" counterfactual experiments for strategic decision-making. Compared with other prediction and experimentation tools such as machine learning and causal graphs, simulation-based prediction has two major benefits: (1) *DES captures event-driven correlation.* Take a supply chain simulation as an example, by simulating the behavior of manufacturers and customers as well as inventory flows of different products together, we can capture the correlation between different types of flows simultaneously through events rather than estimating the correlation from a pure data-driven approach. For example, if a warehouse is experiencing high backlog due to limited labor resource, then all the products that have inbound flows to this warehouse will be impacted with positive correlation. Such event-driven correlation can be naturally reflected via event-based simulation. (2) *DES ensures statistical consistency.* Since a discrete event simulator can model events and entities at the finest granularity where their interaction happens (i.e., manufacturer-product-warehouse level in supply chain business), such a bottom-up modeling structure will simultaneously output metrics of interest at different granularities. Those metrics will preserve statistical consistency and follow restricted physical dynamics as needed.

In spite of these advantages, the forecast accuracy of simulation-based prediction is one of the main challenges in industrial use-cases. Due to inadequate input accuracy and simulation model fidelity, the raw simulation outputs cannot easily meet user's accuracy requirements. Overall, the bias and variance of simulation output metrics arise from three sources: (1) *Monte Carlo (MC) sampling*: Simulating a finite number of replications with different random seeds in each replication introduces MC variance to simulation output. (2) *Input error*: The estimated input distributions from real-world observations could introduce both bias and variance to simulation output. The variance caused by input error is different from MC variance, which we call input uncertainty (IU) (Barton 2012). (3) *Model fidelity gaps*: This is caused by aspects of the system physics (e.g., supply chain system) that are either not modeled in simulation or only modeled with inadequate fidelity in simulation. Model fidelity gaps bring bias to the mean estimators. Taken collectively, we refer to the bias and variance caused by input error and fidelity gaps as *model risk*. Compared with MC variance, model risk often dominates the total risk, especially when the number of replications is large. MC variance is the estimation error for the mean predictions, which can in principle be reduced by running more replications. Model risk, on the other hand, is much more difficult to quantify and cannot be reduced by increasing simulation effort. It can only be reduced by improving simulation input accuracy and model fidelity.

In this paper, we propose INQUIRE, a novel *INput-aware Quantification of Uncertainty for Interpretation, Risk, and Experimentation*. The proposed framework leverages machine learning to predict simulation output uncertainty using key input parameters. The incorporation of input allows users to interpret changes in uncertainty and identify input drivers directly from the same model used for predicting uncertainty; it also supports users in evaluating changes in uncertainty/risk under different policy configurations (i.e., input features), which creates a powerful tool for business users to make risk-aware strategic decisions.

1.2 Literature Review

Uncertainty quantification (UQ) is not a new research topic in stochastic simulation, and there exists a large amount of research work including both frequentist and Bayesian approaches in the simulation literature (Nelson 2010; Riedmaier et al. 2021; Zhu et al. 2020). Most UQ work assumes that simulation output uncertainty is contributed by MC variance and IU, where IU for quantifying and explaining simulation output uncertainty is also well studied in the literature. IU research initially focused on direct resampling methods (i.e., boostrapping) to characterize the error due to input models being fit with finite real-world data. Given the computational burden of boostrapping, Bayesian model averaging (BMA) strategies were developed to characterize IU based on uncertainty in both the input distribution family as well as the input distribution's parameter values (Chick 1997; Chick 1999; Chick 2000; Chick 2001). Another approach to quantifying input uncertainty employs metamodeling techniques, where simulation runs are replaced by a metamodel to minimize simulation error. When using these metamodels (typically Gaussian distributions to represent input parameter uncertainty), the output distribution can be characterized through analytical methods (Ankenman et al. 2010; Barton et al. 2010). However, these methods for identifying the input models that contribute the most to input uncertainty require a sequence of additional diagnostic experiments; in the worst case this requires as many experiments as there are input models, and each of these experiments can be substantial. As an improvement, (Song and Nelson 2015) provides a new analysis that requires only one diagnostic experiment to assess the overall effect of IU, the relative contribution of each input distribution, and a measure of sample size sensitivity of each distribution. Inspired by simulation analytics, (Lin et al. 2015) provides the first single-run method for IU by retaining the simulation sample paths to further reduce computational complexity and deriving the measure of IU variance – both overall variance and the contribution to it of each input model – from the nominal experiment that the analyst would typically run using the estimated input models.

Although there exists a rich amount of research in UQ and IU, much of this work cannot be used directly in large-scale industrial simulations. In practice, running simulations can be very expensive, and the input data dimension is generally high; any methods requiring repeated experiments are therefore not

suitable. Additionally, real-world data are often heavy-tailed and sparse, so how to handle outliers with limited data is a challenge in industrial simulation UQ. Furthermore, existing literature often assumes unbiasedness of simulation mean estimators when estimating uncertainty, while the mean estimators are often biased due to inadequate input and model quality in reality. Uncertainty information is critical for business owners as they need to make long-term strategic decisions with unknown factors. In particular, they are interested in the estimation of uncertainty under both nominal and counterfactual (i.e., "what-if") configurations to make better risk-aware decisions. Understanding uncertainty contribution from each simulation input is also important because it can help business owners prioritize resources to reduce the uncertainty from the top drivers. To enhance the robustness of the decision-making process for downstream simulation users, we propose a novel framework for UQ that enables risk assessment, interpretation, and counterfactual experimentation of large-scale simulations in industry. While the proposed framework can be applied to a broad range of industrial use-cases, in the rest of the paper we apply our proposed methods to retail supply chain management (Acimovic and Graves 2015; Acimovic and Farias 2019).

The rest of the paper is structured as follows: Section 2 introduces the proposed methodology and algorithm for UQ. Section 3 presents the three applications of INQUIRE – quantile prediction for *risk* assessment, uncertainty *interpretation*, risk analysis of *experimentation*. Finally, Section 4 concludes the paper and discusses potential future research directions.

2 INPUT-AWARE UNCERTAINTY QUANTIFICATION

We choose weekly inbound arrival units, a key supply chain metric, as the example metric in this paper to help illustrate our methodology. This metric is defined as the number of units of product received by warehouses from external suppliers every week. In principle, the methodology is metric agnostic and can be applied to all other supply chain metrics or other industries. Suppose we run the supply chain simulation on the current week t and this simulation simulates the supply chain inventory flows for the future N weeks. We are interested in estimating the uncertainty of future inbound arrival units Y_{t+n} during target week $t+n$ for each $n = 1, ..., N$.

Our proposed method focuses on UQ, and we consume a separate mean calibration model which leverages business insights to reduce bias in the estimated means of simulation outputs. This calibration model assembles simulation the raw sample mean and time-series prediction, and has exhibited superior accuracy in the prediction of inbound arrival units compared with the raw simulated mean. For simplicity, we denote the inbound arrival units mean prediction obtained from the calibration approach as calibrated mean henceforth. In this work, we utilize calibrated mean as the mean of our predicted distribution and derive quantile predictions based on it. The reasons for doing so are twofold: first, an accurate mean prediction can significantly improve variance and quantile predictions, especially when the underlying distribution is nearly symmetric; and second, using the calibrated mean ensures that our predicted distributions are consistent with various business systems that have already adopted the mean calibration model.

2.1 Residual-based Uncertainty Prediction

In this subsection, we propose a residual-based uncertainty prediction method, which facilitates the input-aware UQ. Define $\mathscr{F} := \{F_l\}_{l \in \mathscr{L}}$ as the collection of real-world distributions correlated with actual inbound arrival units, e.g., unknown underlying distributions of customer demand, manufacturer lead time (MLT), etc. Suppose the actual inbound arrival units follows $Y(\mathscr{F}) = \mu(\mathscr{F}) + \varepsilon(\mathscr{F})$, where $\mu(\mathscr{F}) = \mathbb{E}[Y(\mathscr{F})]$ is the mean function of the actual inbound arrival units whose input is the set of input distributions $\mathscr{F}$, and $\varepsilon(\mathscr{F}) \sim \mathscr{D}(0, \sigma^2(\mathscr{F}))$ with some distribution $\mathscr{D}$ whose variance also depends on $\mathscr{F}$. We denote $\hat{\mu}(\hat{\mathscr{F}})$ as a predictor of $\mu(\mathscr{F})$, where $\hat{\mathscr{F}} := \{\hat{F}_l\}_{l \in \mathscr{L}'}$ is a set of fitted input distributions $\hat{F}_l$ for estimating F_l for $l \in \mathscr{L}'$, e.g., estimated customer demand and MLT distributions. Note that the number of fitted input distributions $|\mathscr{L}'|$ is not necessary equal to $|\mathscr{L}|$, since some of the distributions influencing the actual inbound arrival units may be unobservable or challenging to estimate accurately.

We are aiming at estimating $\mathrm{Var}(Y(\mathscr{F}) \mid \hat{\mathscr{F}})$, i.e., the variance of the actual inbound arrival units given the fitted input distributions $\hat{\mathscr{F}}$. Estimating the conditional variance poses a significant challenge due to the limited availability of observations and high dimensionality of $\hat{\mathscr{F}}$. At any given week, we can only obtain a single observation, making it impossible to calculate the sample variance and fit a model directly for variance prediction. Consequently, we employ a residual-based model as an alternative approach to forecasting the variance. To understand how this approach works out, we first note that

$$\mathrm{Var}\left(Y(\mathscr{F}) \mid \hat{\mathscr{F}}\right) = \mathrm{Var}\left(\varepsilon(\mathscr{F}) \mid \hat{\mathscr{F}}\right) = \mathrm{Var}\left(Y(\mathscr{F}) - \mu(\mathscr{F}) \mid \hat{\mathscr{F}}\right) = \mathbb{E}\left[(Y(\mathscr{F}) - \mu(\mathscr{F}))^2 \mid \hat{\mathscr{F}}\right].$$

Thus, one way to predict $\mathrm{Var}(Y(\mathscr{F}) \mid \hat{\mathscr{F}})$ is to predict the expectation of $(Y(\mathscr{F}) - \mu(\mathscr{F}))^2$ conditioning on $\hat{\mathscr{F}}$. Since $\mu(\mathscr{F})$ is unattainable in practice, we replace it by $\hat{\mu}(\hat{\mathscr{F}})$ and calculate the squared residual as

$$r(\hat{\mathscr{F}}) = (Y(\mathscr{F}) - \hat{\mu}(\hat{\mathscr{F}}))^2.$$

Using the realizations of $r(\hat{\mathscr{F}})$ derived from the historical observations of $Y(\mathscr{F})$, we can construct a prediction model $\hat{f}(\hat{\mathscr{F}})$ to predict $\mathbb{E}[r(\hat{\mathscr{F}}) \mid \hat{\mathscr{F}}]$ based on observed $r(\hat{\mathscr{F}})$ and input samples from $\hat{\mathscr{F}}$. Note that replacing $\mu(\mathscr{F})$ by $\hat{\mu}(\hat{\mathscr{F}})$ introduces bias into the prediction of $\mathrm{Var}\left(\varepsilon(\mathscr{F}) \mid \hat{\mathscr{F}}\right)$. To demonstrate it, consider the following decomposition:

$$\begin{aligned}\mathbb{E}\left[r(\hat{\mathscr{F}}) \mid \hat{\mathscr{F}}\right] =& \mathbb{E}\left[(Y(\mathscr{F}) - \hat{\mu}(\hat{\mathscr{F}}))^2 \mid \hat{\mathscr{F}}\right] = \mathbb{E}\left[(Y(\mathscr{F}) - \mu(\mathscr{F}) + \mu(\mathscr{F}) - \hat{\mu}(\hat{\mathscr{F}}))^2 \mid \hat{\mathscr{F}}\right] \\ =& \mathrm{Var}\left(\varepsilon(\mathscr{F}) \mid \hat{\mathscr{F}}\right) + \left(\mathbb{E}\left[\mu(\mathscr{F}) - \hat{\mu}(\hat{\mathscr{F}}) \mid \hat{\mathscr{F}}\right]\right)^2 + \mathrm{Var}\left(\hat{\mu}(\hat{\mathscr{F}}) \mid \hat{\mathscr{F}}\right) \end{aligned} \tag{1}$$

The first term is equal to $\mathrm{Var}(Y(\mathscr{F}) \mid \hat{\mathscr{F}})$, which is our objective. The second term is the fidelity gap introduced by the bias of the mean prediction and the input uncertainty from the input distribution predictions. The third term is the uncertainty of the mean prediction, which depends on the properties of the mean prediction model. If $\hat{\mu}(\hat{\mathscr{F}})$ is the sample mean of the simulation outputs using the fitted input distribution $\hat{\mathscr{F}}$ and M is the number of samples taken from the simulation, then $\mathrm{Var}(\hat{\mu}(\hat{\mathscr{F}}) \mid \hat{\mathscr{F}}) = M^{-1}\mathrm{Var}(\varepsilon(\hat{\mathscr{F}}) \mid \hat{\mathscr{F}})$ is the Monte Carlo simulation error, which diminishes as $M \to \infty$. If $\hat{\mu}(\hat{\mathscr{F}})$ is a prediction model built on the historical samples means with corresponding inputs, e.g., calibrated mean, then $\mathrm{Var}(\hat{\mu}(\hat{\mathscr{F}}) \mid \hat{\mathscr{F}})$ accounts for both the randomness of the training data set and the training procedure.

As (1) shows, to predict $\mathrm{Var}(Y(\mathscr{F}) \mid \hat{\mathscr{F}})$, suppose $\hat{f}(\hat{\mathscr{F}})$ is an unbiased predictor of $\mathbb{E}[r(\hat{\mathscr{F}}) \mid \hat{\mathscr{F}}]$, a debiased predictor

$$\hat{\sigma}^2(\hat{\mathscr{F}}) = \hat{f}(\hat{\mathscr{F}}) - \left(\mathbb{E}\left[\mu(\mathscr{F}) - \hat{\mu}(\hat{\mathscr{F}}) \mid \hat{\mathscr{F}}\right]\right)^2 - \mathrm{Var}\left(\hat{\mu}(\hat{\mathscr{F}}) \mid \hat{\mathscr{F}}\right)$$

can serve as a predictor of $\mathrm{Var}(Y(\mathscr{F}))$. However, since both the second and the third terms on the right hand side (r.h.s.) are difficult to eliminate in practice, we use $\hat{f}(\hat{\mathscr{F}})$ instead of $\hat{\sigma}^2(\hat{\mathscr{F}})$ as the variance predictor, accounting for uncertainty from all sources listed above. In fact, if the calibrated mean model is accurate enough, i.e., $\mu(\mathscr{F}) \approx \mu(\hat{\mathscr{F}})$, then the second and third terms are negligible such that $\hat{\sigma}^2(\hat{\mathscr{F}}) \approx \hat{f}(\hat{\mathscr{F}})$.

To construct the uncertainty prediction model $\hat{f}(\hat{\mathscr{F}})$, we can leverage machine learning models to learn the relationship between observed $r(\hat{\mathscr{F}})$ and input samples from $\hat{\mathscr{F}}$. The choice of model depends on data characteristics and the trade-offs between prediction power and computation speed. We offer more relevant discussions in Section 3.1.1. In the model training process, since both $Y(\mathscr{F})$ and $\hat{\mu}(\hat{\mathscr{F}})$ are realized, we can scale $r(\hat{\mathscr{F}})$ by $(\hat{\mu}(\hat{\mathscr{F}}))^2$, i.e.,

$$\tilde{r}(\hat{\mathscr{F}}) = r(\hat{\mathscr{F}})/(\hat{\mu}(\hat{\mathscr{F}}))^2 = (Y(\mathscr{F})/\hat{\mu}(\hat{\mathscr{F}}) - 1)^2, \tag{2}$$

and use the scaled residual $\tilde{r}(\hat{\mathscr{F}})$ as the observations to construct the prediction model. The predicted variance of inbound arrival units can be recovered by multiplying the mean prediction $\hat{\mu}^2(\hat{\mathscr{F}})$ at the target week. There are two main reasons for adopting this ratio-based approach. First, it can mitigate the impact of outliers. Although the absolute prediction errors may vary significantly across different periods, the relative errors can be similar. Consequently, predicting the variance of the ratio tends to be more accurate and easier. Second, previous UQ studies have demonstrated the superior performance of the ratio-based form. Our numerical tests also corroborate its advantages over directly estimating the variance of the original random variable. By utilizing the predicted variance and the calibrated mean, we can construct the predicted distributions by assuming a specific distribution family.

3 APPLICATIONS OF INQUIRE

We apply the input-aware UQ framework to three applications: quantile prediction for risk assessment, uncertainty interpretation, and counterfactual experimentation. We present the additional algorithms needed for each application, and illustrate the effectiveness via numerical examples.

3.1 Quantile Prediction and Prediction Interval Construction for Risk Assessment

The objective is to construct an accurate prediction $\widehat{Y}^q_{t+n}$ of the q-quantile of Y_{t+n}, such that $\mathbb{P}(Y_{t+n} \leq \widehat{Y}^q_{t+n}) = q$. Additionally, we aim to construct a *prediction interval* (PI) centered around mean, $\widehat{PI}^q_{t+n}$, with the desired q-coverage probability, i.e., $\mathbb{P}(Y_{t+n} \in \widehat{PI}^q_{t+n}) = q$.

Predicting quantiles and constructing PIs with assumptions on normality may yield unsatisfactory results, particularly when the underlying distribution is asymmetric. As demonstrated later in our numerical experiments, when the predicted 0.5 quantile (given by the estimated mean) exhibits over- or under-coverage, the overall coverage of predicted quantiles from 0.1 to 0.9 can be severely affected. To address this issue, we propose a skewness-detection method. The main idea is to first estimate the nonparametric skewness of the predicted distribution. If the nonparametric skew is significantly different from 0, indicating an asymmetric distribution, we construct the predicted quantiles and PIs using a skewed-Gaussian distribution. Otherwise, we revert to the normal distribution assumption. Note that this approach relies strongly on the belief that the predicted mean we adopted (calibrated mean in our case) has negligible bias; if this is not the case, then the skewness can be misdetected, leading to an even worse performance than that without skewness detection.

The main procedure of the skewness-detection is shown in Algorithm 1. The training window is denoted by T. For $t = 1, 2, \ldots, T$, x_t represents the observed input vector at week t from $\hat{\mathscr{F}}$, comprising quantile predictions of demand and MLT, for instance. Additionally, y_t denotes the corresponding actual inbound arrival units. The parameter δ in line 8 is a user-specified threshold that governs the choice between utilizing a skewed-Gaussian distribution or Gaussian distribution. The selection of δ can be based on empirical evidence or determined through cross-validation techniques. Typically, a larger value of δ results in a more conservative decision regarding the application of a skewed-Gaussian distribution. As $\delta \to \infty$, the method reverts to the scenario where a Gaussian distribution is assumed. In our implementation, we choose $\delta = 0.05$ by empirical evaluations. With estimated parameters, we can estimate the mean and variance of a skewed-Gaussian distribution (Azzalini and Arellano-Valle 2013; Hou et al. 2021).

3.1.1 A Numerical Illustration

We apply Algorithm 1 for inbound arrival units UQ, adopting the XGBoost library (Chen and Guestrin 2016) for variance prediction and quantile regression forests (Meinshausen and Ridgeway 2006) for median prediction. We opted for tree-based models as they demonstrate better prediction power under limited data and offer superior interpretability compared to deep learning models, which are crucial advantages for our problems given the limited actual observations we have and user's need for uncertainty interpretation. Additionally, our own numerical study validates the effectiveness of tree-based models, outperforming other traditional methods such as linear regression and Gaussian Process regression. We use Amazon retail

Algorithm 1: Skew-detection algorithm for quantile prediction PI construction

Input: Training data set $\mathscr{M} = \{(x_t, y_t)\}_{t=1}^{T}$, predicted means $\{\hat{\mu}_t\}_{t\geq 1}$, skewness tolerance δ
Output: Predicted quantiles and PIs

```
Compute the errors r_t = (y_t − μ̂_t)^2 for t = 1, 2, ..., T;
Construct training data set R = {(x_t, r_t)}_{t=1}^T;
Fit the prediction model f̂ of variance using training data set R;
Fit the prediction model ĝ of median using training data set M;
Predict the median m_t = ĝ(x_t) and variance v_t = f̂(x_t);
if |m_t − μ̂_t|/√v_t > δ then
    Fit a skew-normal distribution;
    Construct predicted quantiles and PIs using the skew-normal distribution;
else
    Construct predicted quantiles and PIs using normal distribution;
```

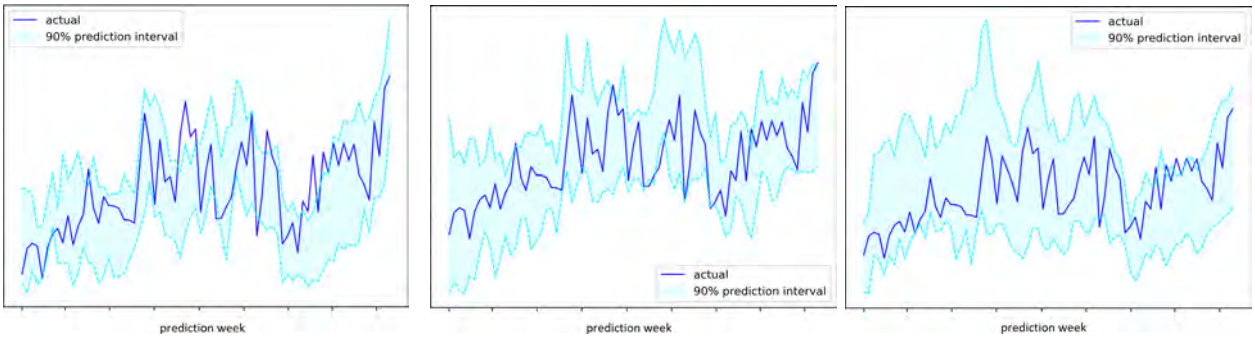
supply chain data in our example, specifically, the inputs we choose include inventory, purchased orders from manufacturer, key quantiles of customer demand and MLT forecasts, economic cost, and historical ratios of actual observation over forecast. We use historical two-year weekly inbound arrival units and the corresponding inputs as the training data, and test the performance of predicted quantiles and PIs using the next one-year data. For comparison, we include the following four methods: (1) assuming a Gaussian distribution for uncertainty prediction, estimating the variance of the distribution via maximum likelihood estimation (MLE), and centering the distribution on the raw simulated arrivals mean (we referred to as MLE-raw); (2) applying the same Gaussian MLE approach in (1) to estimate uncertainty but centering the distribution around the calibrated mean simulation output (MLE-calibrated); (3) predicting the variance of a Gaussian using XGBoost and centering the distribution on the calibrated mean simulation output (XG-calibrated); and (4) adding skewness-detection to the XG-calibrated model (XG-SD-calibrated). To measure the overall performance, we define the following metrics:

- Absolute error (AE) of empirical coverage: Letting $\hat{q}$ denote the empirical coverage for a target coverage $q \in \mathscr{Q}$, the AE is calculated as $|\mathscr{Q}|^{-1}\sum_{q\in\mathscr{Q}}|\hat{q}-q|$.
- Continuous ranked probability score (CRPS): Define the quantile loss at q as $\mathrm{QL}_q(\hat{Y}^q, Y) = Y^{-1}(q\cdot(Y-\hat{Y}^q)^+ + (1-q)\cdot(\hat{Y}^q - Y)^+)$, where $\hat{Y}^q$ is the q-quantile prediction. Denote the set of inbound arrival observations as $\mathscr{D}$. The CRPS is then calculated as $2\sum_{q\in Q}|Y|^{-1}\sum_{y\in\mathscr{D}}\mathrm{QL}_q(\hat{Y}^q, y)$.

Table 1 shows the empirical coverage of the quantile predictions of week $t+6$, i.e., the 6-weeks-ahead forecast. As we can see, the calibrated mean does not serve as an accurate median prediction under the normal distribution assumption, resulting in the over-coverage for the quantile predictions of p30 and p40. Adopting the skew-detection algorithm successfully detects this phenomenon, and the adjusted asymmetric distribution gives a more accurate quantile prediction from both AE and CRPS. Figures 1 shows the PIs with 90% coverage (i.e., 0.95-quantile prediction − 0.05-quantile prediction) of week $t+1$, $t+3$ and $t+6$, which indicate the method provides satisfactory coverage.

Table 1: Coverage of the quantiles for week $t+6$.

Methods	p10	p20	p30	p40	p50	p60	p70	p80	p90	AE	CRPS
MLE-raw	8.22%	27.40%	49.32%	58.90%	69.86%	78.08%	87.67%	90.41%	93.15%	12.95%	0.1152
MLE-calibrated	4.11%	19.18%	36.99%	50.68%	57.53%	61.64%	65.75%	78.08%	89.04%	4.52%	0.0932
XG-calibrated	13.70%	24.66%	36.99%	47.95%	57.53%	61.64%	67.12%	76.71%	91.78%	4.49%	0.0942
XG-SD-calibrated	9.59%	21.92%	28.77%	39.73%	50.68%	56.16%	68.49%	76.71%	87.67%	1.72%	0.0898



(a) Prediction of week $t+1$. (b) Prediction of week $t+3$. (c) Prediction of week $t+6$.

Figure 1: Actual inbound arrival units versus mean and 90% PIs. Exact prediction values in the y-axis are omitted due to the company's confidentiality policy.

3.2 Uncertainty Interpretation of the Predicted Uncertainty

INQUIRE also enables efficient uncertainty interpretation, which is our second application in this paper. As mentioned in Section 1, the relative contribution from each driver to the overall uncertainty is important because it helps business owners prioritize resources to reduce the uncertainty from top drivers. We employ the Shapley value, which is defined in (3) in the following section, to explain simulation output uncertainty, and we aim at answering two questions: (1) Which inputs drive uncertainty change for two simulations, and (2) what are the top drivers of uncertainty for a single simulation run? The key idea of Shapley value is for a given input i, compare the average output difference for those outputs produced by coalitions that include input i verses those that do not (Owen 2014; Lundberg and Lee 2017; Burkart and Huber 2021). In other words, for each input i we first consider all input coalitions (different subsets of the inputs) with i, replace the values with all other possible values in these subsets, and average their corresponding outputs. Then we repeat for all coalitions without input i. The attribution to input i is then given by the difference of these average outputs. When the n inputs are independent and the functional relationship between inputs and output is linear, Shapley value simplifies to the product of the coefficient (estimated by a linear regression) and the input change value. When the functional relationship is non-linear, if using a black-box model, computing Shapley values usually requires Monte Carlo estimation where multiple subsets of the corresponding inputs are substituted with different values through sampling (Song et al. 2016).

3.2.1 Shapley Value

Definition 1 (Shapley value) For a D-player cooperative game with the set of players $\mathscr{D} = \{1,2,\ldots,D\}$ and gain $c(\cdot)$, the Shapley value of the i^{th} player is defined by:

$$\phi_i = \sum_{\pi \in \Pi(\mathscr{D})} \frac{1}{D!} \left(c(P_i(\pi) \cup \{i\}) - c(P_i(\pi)) \right), \tag{3}$$

where $\Pi(\mathscr{D})$ denotes the set of all $D!$ permutations of players in $\mathscr{D}$, and $P_i(\pi)$ denotes the players that precede player i in one permutation π.

First introduced in (Shapley 1953), Shapley value has many convenient properties such as efficiency, symmetry, dummy and additivity. These properties are critical for our interpretation problem, e.g., the efficiency property ensures that the sum of the contribution value to each player (which are the inputs in simulation) is equal to the total uncertainty change from two simulation runs (which is the target difference to explain). The Shapley Value has been applied in different areas such as game theory and machine learning, and we adopt its application in machine learning in this paper. In machine learning prediction, the Shapley value is used to explain the contribution of each input to the difference of a prediction and the average.

For example, consider a machine learning model that predicts apartment prices: For a certain apartment it predicts \$300,000. The average prediction for all apartments is \$310,000. Then the Shapley value is used to explain how much has each input value contributed to this certain prediction (\$310,000) compared to the average prediction (\$300,000)? Let this machine learning function used to predict apartment price to be denoted as $p(\mathbf{x})$, then according to efficiency property, the sum of the Shapley values (input contribution) of all players (inputs) must equal the difference of prediction for $\mathbf{x}$ and the expectation, which is taken with respect to (w.r.t.) the sampling noise when drawing $\mathbf{x}$ from $\mathscr{F}_{\mathbf{X}}$.

$$\sum_{\pi \in \Pi(\mathscr{D})} \phi_i = p(\mathbf{x}) - \mathbb{E}[p(\mathbf{X})]. \tag{4}$$

3.2.2 Shapley Value Based Uncertainty Interpretation

Continued with the notation defined in Section 2.1, assume we have predicted simulation output uncertainty $\hat{f}_j(\hat{\mathscr{F}}_j)$ from the jth run, $j = 0, 1, \ldots, J-1$, and let the Shapley value for each input i from the jth simulation run be ϕ_i^{ℓ}. Suppose we are interested in explaining the uncertainty change from specific two runs, $\hat{f}_j(\hat{\mathscr{F}}_j)$ vs $\hat{f}_s(\hat{\mathscr{F}}_s)$ where $j \neq s$, then according to (4), the Shapley values computed from these two simulation runs satisfy

$$\sum_{l \in \mathscr{L}} \phi_{l,j} = \hat{f}_j(\hat{\mathscr{F}}_j) - \mathbb{E}(\hat{f}(\hat{\mathscr{F}})) \quad \text{and} \quad \sum_{l \in \mathscr{L}} \phi_{l,s} = \hat{f}_s(\hat{\mathscr{F}}_s) - \mathbb{E}(\hat{f}(\hat{\mathscr{F}})),$$

where the expectation here is taken with respect to the randomness in $\hat{\mathscr{F}}$. Recall $\hat{\mathscr{F}}$ is estimated input distributions from real-world data and the underlying true input $\mathscr{F}$ is unknown, so there exists uncertainty in estimated $\hat{\mathscr{F}}$. We are interested in explaining the contribution of each input to the variance difference, i.e., $\hat{f}_j(\hat{\mathscr{F}}_j) - \hat{f}_s(\hat{\mathscr{F}}_s)$, therefore we take the difference of the two equations above to get

$$\sum_{l \in \mathscr{L}} \phi_{l,j} - \sum_{l \in \mathscr{L}} \phi_{l,s} = \sum_{l \in \mathscr{L}} (\underbrace{\phi_{l,j} - \phi_{l,s}}_{\Phi_l}) = \hat{f}_j(\hat{\mathscr{F}}_j) - \hat{f}_s(\hat{\mathscr{F}}_s) \tag{5}$$

where Φ_l, the difference of the two Shapley values, is the variance attribution of input $\hat{\mathscr{F}}_l$.

Typically, the interpretation of the change of the predictions requires us using the same variance prediction model $\hat{f}$. However, we could build separate prediction models for different target weeks due to business requirements, and using separate prediction models in the interpretation procedure can lead to consistency issues. Additionally, note that because the prediction model predicts the variance of the ratio which is defined in (2), we need to multiply it by the squared calibrated mean $(\hat{\mu}(\hat{\mathscr{F}}))^2$ to get the predicted variance of the inbound arrival units. Since the calibrated mean prediction for a given target week can differ across adjacent simulation weeks t and $t' \neq t$, it introduces another source of inconsistency. To illustrate, consider two prediction tasks involving simulations run j and s on weeks t and $t+1$, with both simulations predicting the same target week $t+1$. Denoting the predicted variance of the ratios from the week t and week $t+1$ simulations as $\hat{v}_j(\hat{\mathscr{F}}_j)$ and $\hat{v}_s(\hat{\mathscr{F}}_s)$, respectively, (5) can be written as

$$\sum_{l \in \mathscr{L}} \phi_{l,j} - \sum_{l \in \mathscr{L}} \phi_{l,s} = \hat{v}_j(\hat{\mathscr{F}}_j) - \hat{v}_s(\hat{\mathscr{F}}_s),$$

if the simulation means are identical. However, given the calibrated predictions $\hat{\mu}_j(\hat{\mathscr{F}}_j) \neq \hat{\mu}_s(\hat{\mathscr{F}}_s)$, we have

$$\hat{f}_j(\hat{\mathscr{F}}_j) - \hat{f}_s(\hat{\mathscr{F}}_s) = \hat{\mu}_j^2(\hat{\mathscr{F}}_j)\hat{v}_j(\hat{\mathscr{F}}_j) - \hat{\mu}_s^2(\hat{\mathscr{F}}_s)\hat{v}_s(\hat{\mathscr{F}}_s) \neq \hat{\mu}_j^2(\hat{\mathscr{F}}_j) \cdot \sum_{l \in \mathscr{L}} \phi_{l,j} - \hat{\mu}_s^2(\hat{\mathscr{F}}_s) \cdot \sum_{l \in \mathscr{L}} \phi_{l,s},$$

which violates the efficiency property of the Shapley value. To fix this issue, we first construct a unified variance prediction model $\hat{g}$, which utilizes all training samples for predicting week t and $t+1$, and do the prediction for the same target week $t+1$. The Shapley value derived from $\hat{g}$ should maintain its efficiency

property. Then, we use two separate prediction models for prediction tasks. To attribute the predicted variance change, i.e., $\hat{f}_j(\hat{\mathscr{F}}_j) - \hat{f}_s(\hat{\mathscr{F}}_s)$, we use the following scaling method:

$$\Phi'_l = (\phi_{l,j} - \phi_{l,s}) \cdot \frac{\hat{f}_j(\hat{\mathscr{F}}_j) - \hat{f}_s(\hat{\mathscr{F}}_s)}{\hat{v}_j(\hat{\mathscr{F}}_j) - \hat{v}_s(\hat{\mathscr{F}}_s)}, \quad \forall l \in \mathscr{L}, \tag{6}$$

where Φ'_l is the scaled Shapley value explaining the contribution of the change of each input to the change of the predicted variance. The scaled Shapley values preserve the efficiency property.

Our framework can also quantify the contribution of each input to the total uncertainty in a single run. Unlike explaining the uncertainty change from two runs, in which the efficiency property of the Shapley values may be destroyed, as discussed above, the single-week attribution procedure is straightforward. We take run j as an example. Based on (4), we know $\sum_{l\in\mathscr{L}} \phi_{l,j} = \hat{v}_j(\hat{\mathscr{F}}_j) - \mathbb{E}(\hat{v}(\hat{\mathscr{F}}))$, let scaling factor $\alpha = \hat{v}_j(\hat{\mathscr{F}}_j)/(\hat{v}_j(\hat{\mathscr{F}}_j) - \mathbb{E}(\hat{v}(\hat{\mathscr{F}})))$, we have

$$\phi'_{l,j} = \alpha \phi_{l,j} \cdot \hat{\mu}_j^2(\hat{\mathscr{F}}_j), \quad \forall l \in \mathscr{L},$$

The scaled Shapley values quantify the contribution of inputs to the predictive variance $\hat{f}_j(\hat{\mathscr{F}}_j)$:

$$\sum_{l\in\mathscr{F}} \phi'_{l,j} = \alpha \sum_{l\in\mathscr{L}} \phi_{l,j} \cdot \hat{\mu}_j^2(\hat{\mathscr{F}}_j) = \alpha \left(\hat{v}_j(\hat{\mathscr{F}}_j) - \mathbb{E}(\hat{v}(\hat{\mathscr{F}}))\right) \cdot \hat{\mu}_j^2(\hat{\mathscr{F}}_j) = \hat{v}_j(\hat{\mathscr{F}}_j) \cdot \hat{\mu}_j^2(\hat{\mathscr{F}}_j) = \hat{f}_j(\hat{\mathscr{F}}_j).$$

3.2.3 Numerical Results

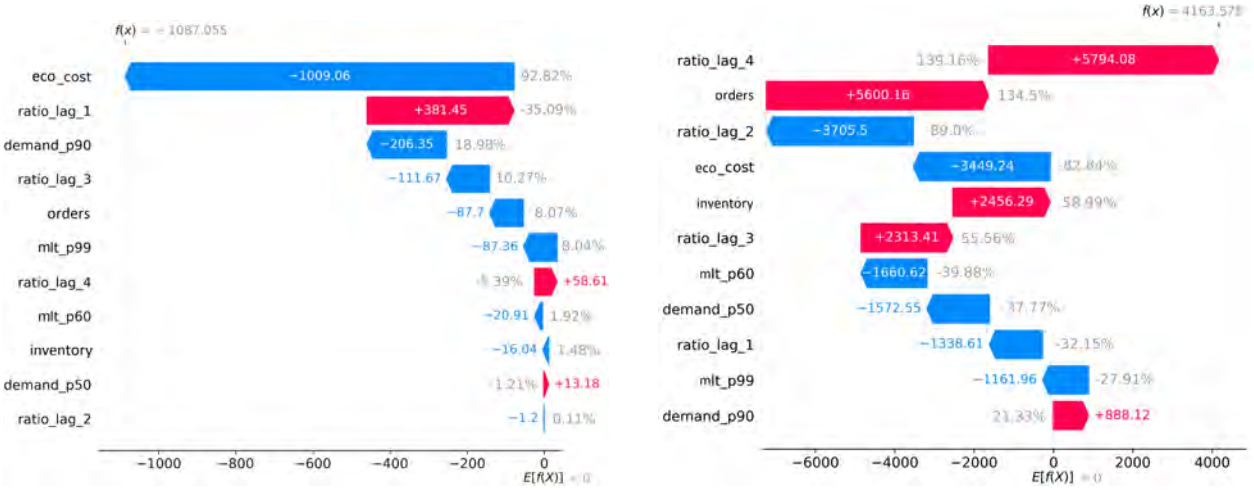
We choose the same inputs used in quantile prediction example (Section 3.1.1) for interpretation in this section. Similarly, we continue with XGBoost which provides efficient Shapley value computation in addition to its superior prediction power. Figure 2 (a) shows the interpretation of predicted uncertainty change from two simulation runs. As we can see, the economic cost is the main driver for the decrease of the uncertainty, which aligns with the decrease in its raw value. Additionally, "ratio lag 1", which is the ratio from previous week's forecast and actual observation, contributed negatively to the uncertainty reduction. We notice the value of ratio lag 1 drops from 0.8 to 0.69, i.e., the historical accuracy of the calibrated mean is potentially getting worse. Thus, the predicted simulation model fidelity gap increases, leading to the increase of the predicted variance. Figure 2 (b) shows the contribution of inputs to the relative total uncertainty of a single target week.

3.3 Counterfactual Experimentation

The last application of INQUIRE is to support business counterfactual experimentation. Incorporating inputs that are highly correlated with the supply chain system into our proposed UQ model enables counterfactual analysis, which is critical for business owners to make strategic decisions. Counterfactual analysis poses a significant challenge because the ground truth of the counterfactual scenario is never known, making it difficult to evaluate the proposed method's capability for such analysis. In this section, we first consider a scenario under different hypothetical economic cost, as we can potentially estimate the ground truth at certain cost to evaluate the performance of our UQ approach. Subsequently, we conduct the analysis of scenarios with different economic costs.

We start with the approach for constructing the "ground truth" for scenario where the actual observation cannot be observed. We adopt the notations defined in Section 2 but simplify $Y(\mathscr{F})$ to Y and $\hat{\mu}(\mathscr{F})$ to $\hat{\mu}$ for simplicity. Let Y_u and Y_c denote the inbound arrival units under unobservable and observable scenarios, respectively, and $\hat{\mu}_u^{\text{sim}}$ and $\hat{\mu}_c^{\text{sim}}$ represent the raw simulated arrivals. Recall that our primary goal is to predict the inbound arrival units uncertainty under unobservable scenario, i.e.,

$$\text{Var}(Y_u) = \text{Var}\left(\frac{Y_u}{\hat{\mu}_u^{\text{sim}}} \cdot \hat{\mu}_u^{\text{sim}}\right).$$



(a) Interpretation of predicted uncertainty change for two simulation runs.

(b) Interpretation of predicted uncertainty of a single run.

Figure 2: Uncertainty interpretation represented by Shapley values (scaled by 10^{10}). The percentage presents the relative contribution of the inputs. The left figure shows the interpretation of changes from two runs, and the right shows the contribution of a single run.

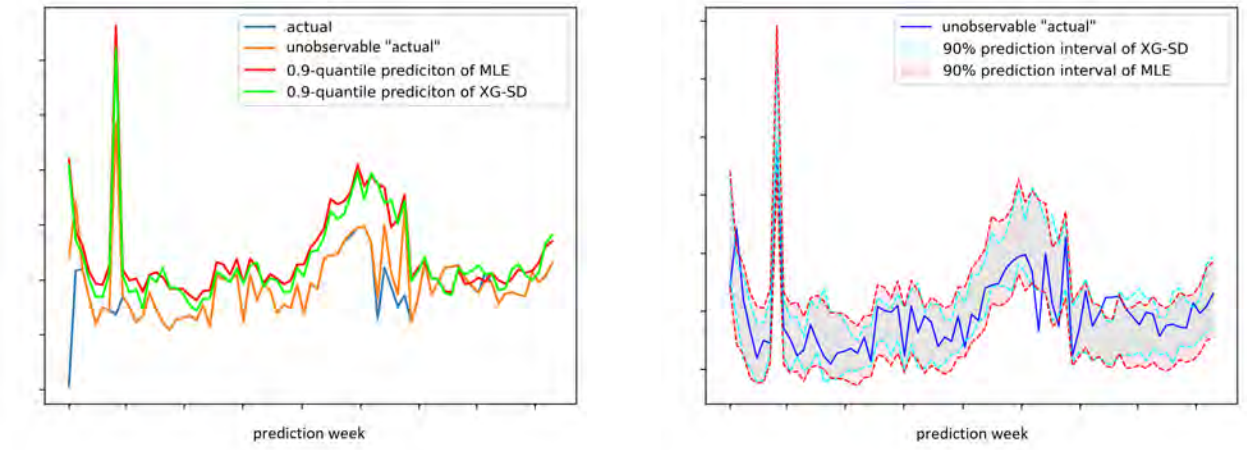
The estimation of inbound arrival units variance under an unobservable scenario is based on our view of $Y_u/\hat{\mu}_u^{\text{sim}}$. Given that Y_u cannot be observed, we need to make assumptions in order to conduct the counterfactual analysis. Specifically, we assume that for each week, the ratio $Y_u/\hat{\mu}_u^{\text{sim}}$ follows the same distribution as $Y_c/\hat{\mu}_c^{\text{sim}}$ at that week, if the actual arrival is observable in that week. Then, the unobservable "actual inbound arrival units" Y_u can be estimated as follows:

$$\hat{Y}_u \approx \frac{Y_c}{\hat{\mu}_c^{\text{sim}}} \cdot \hat{\mu}_u^{\text{sim}}.$$

Although the estimated "actual inbound arrival units" under the unconstrained scenario cannot be fully trusted, it serves as a valuable tool to evaluate the performance of the quantile predictions and predicted intervals. To predict the variance of Y_u, we employ the same inputs as those used during the observable period, with the exception that the economic cost is set to a new hypothetical cost. After deriving the predicted variance of the ratio using the algorithm discussed in Section 2.1, we multiply it by the unconstrained calibrated mean to obtain the prediction of $\text{Var}(Y_u)$. We further evaluate the benchmark method MLE, where the predicted variance of the ratio is unaffected by changes in the economic cost. This evaluation is designed to explore the capability of the ratio-based method for counterfactual analysis when the desired counterfactual input is not included in the prediction model.

Figure 3 shows the quantile prediction and the PIs with 90% coverage of both methods, and Table 2 presents the coverage statistics. We set the hypothetical economic cost to zero in this example, where the actual economic cost is nonzero. The results illustrate that our proposed method offers satisfactory coverage for both quantile predictions and PIs. And MLE-calibrated tends to yield larger quantile predictions and wider PIs in most cases. However, it retains the ability to respond to cost changes thanks to its ratio-based approach. Thus, under our assumption, even if the input is not incorporated into the variance prediction model, we can still perform the counterfactual analysis if the mean prediction model has the capability.

We propose the following general approach for counterfactual analysis. If the counterfactual scenario involves inputs already incorporated in our variance prediction model, we suggest employing our proposed method for analysis. However, if the counterfactual scenario involves inputs not accounted for in our model, we recommend adopting the MLE approach due to its simplicity and ease of implementation.



(a) Quantile prediction.

(b) PI.

Figure 3: Unobservable "actual inbound arrival units" versus 0.9-quantile predictions (left) and 90% coverage PIs (right) for prediction target week $t+1$.

Table 2: Coverage of the quantiles and PIs for week $t+1$ under unobservable scenario.

quantile prediction	p10	p20	p30	p40	p50	p60	p70	p80	p90	AE
XG-SD-calibrated	15.07%	23.29%	30.14%	45.21%	49.32%	54.79%	69.86%	82.19%	90.41%	2.48%
MLE-calibrated	4.11%	17.81%	27.4%	38.36%	49.32%	60.27%	73.97%	83.56%	91.78%	2.51%
PI	p10	p20	p30	p40	p50	p60	p70	p80	p90	AE
XG-SD-calibrated	5.48%	9.59%	27.40%	39.73%	49.32%	58.90%	65.75%	75.34%	87.67%	3.42%
MLE-calibrated	10.96%	21.92%	38.36%	48.58%	56.16%	65.75%	76.71%	87.67%	93.15%	5.25%

4 CONCLUSION AND FUTURE DIRECTIONS

In this work, we developed INQUIRE, a novel framework that significantly improves uncertainty quantification through better quantile predictions and prediction intervals. We demonstrated its effectiveness in three key applications: risk assessment through skewness-aware quantile prediction, uncertainty interpretation via Shapley values, and counterfactual experimentation for strategic decision-making. Looking ahead, key research directions include: developing a unified framework for interpreting both mean and uncertainty changes, and extending uncertainty attribution to inputs not currently included in our prediction model, while maintaining prediction accuracy.

REFERENCES

Acimovic, J., and V. F. Farias. 2019. "The Fulfillment-Optimization Problem". In *Operations Research & Management Science in the Age of Analytics*, 218–237. INFORMS.

Acimovic, J., and S. C. Graves. 2015. "Making Better Fulfillment Decisions on the Fly in an Online Retail Environment". *Manufacturing & Service Operations Management* 17(1):34–51.

Ankenman, B., B. L. Nelson, and J. Staum. 2010. "Stochastic Kriging for Simulation Metamodeling". *Operations Research* 58(2):371–382.

Azzalini, A., and R. B. Arellano-Valle. 2013. "Maximum Penalized Likelihood Estimation for Skew-Normal and Skew-t Distributions". *Journal of Statistical Planning and Inference* 143(2):419–433.

Barton, R. R. 2012. "Tutorial: Input Uncertainty in Output Analysis". In *2012 Winter Simulation Conference (WSC)* https://doi.org/10.1109/WSC.2012.6465266.

Barton, R. R., B. L. Nelson, and W. Xie. 2010. "A Framework for Input Uncertainty Analysis". In *2010 Winter Simulation Conference (WSC)*, 1189–1198 https://doi.org/10.1109/WSC.2010.5679071.

Burkart, N., and M. F. Huber. 2021. "A Survey on the Explainability of Supervised Machine Learning". *Journal of Artificial Intelligence Research* 70:245–317.

Chen, T., and C. Guestrin. 2016. "Xgboost: A Scalable Tree Boosting System". In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chick, S. E. 1997. "Bayesian Analysis for Simulation Input and Output". In *1997 Winter Simulation Conference (WSC)*, 253–260 https://doi.org/10.1145/268437.268488.

Chick, S. E. 1999. "Steps to Implement Bayesian Input Distribution Selection". In *1999 Winter Simulation Conference (WSC)*, 317–324 https://doi.org/10.1109/WSC.1999.823090.

Chick, S. E. 2000. "Bayesian Methods for Simulation". In *2000 Winter Simulation Conference (WSC)*, 109–118 https://doi.org/10.1109/WSC.2000.899705.

Chick, S. E. 2001. "Input Distribution Selection for Simulation Experiments: Accounting for Input Uncertainty". *Operations Research* 49(5):744–758.

Hou, G., A. Xu, F. Cai, and Y.-G. Wang. 2021. "Parameter Estimation for Univariate Skew-Normal Distribution Based on the Modified Empirical Characteristic Function". *Communications in Statistics-Theory and Methods* 51(22):7897–7910.

Lin, Y., E. Song, and B. L. Nelson. 2015. "Single-experiment Input Uncertainty". *Journal of Simulation* 9(3):249–259.

Lundberg, S. M., and S.-I. Lee. 2017. "A Unified Approach to Interpreting Model Predictions". *Advances in Neural Information Processing Systems* 30.

Meinshausen, N., and G. Ridgeway. 2006. "Quantile Regression Forests". *Journal of Machine Learning Research* 7(6).

Nelson, B. L. 2010. *Stochastic Modeling: Analysis & Simulation*. Courier Corporation.

Owen, A. B. 2014. "Sobol' Indices and Shapley Value". *SIAM/ASA Journal on Uncertainty Quantification* 2(1):245–251.

Riedmaier, S., B. Danquah, B. Schick, and F. Diermeyer. 2021. "Unified Framework and Survey for Model Verification, Validation and Uncertainty Quantification". *Archives of Computational Methods in Engineering* 28:2655–2688.

Shapley, L. S. 1953. *A Value for n-Person Games*. Princeton University Press.

Song, E., and B. L. Nelson. 2015. "Quickly Assessing Contributions to Input Uncertainty". *IIE Transactions* 47(9):893–909.

Song, E., B. L. Nelson, and J. Staum. 2016. "Shapley Effects for Global Sensitivity Analysis: Theory and Computation". *SIAM/ASA Journal on Uncertainty Quantification* 4(1):1060–1083 https://doi.org/10.1137/15M1048070.

Zhu, H., T. Liu, and E. Zhou. 2020. "Risk Quantification in Stochastic Simulation under Input Uncertainty". *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 30(1):1–24.

AUTHOR BIOGRAPHIES

YUJING LIN is a Senior Research Scientist from Supply Chain Optimization Technology team in Amazon. She received her Ph.D. degree in Industrial Engineering and Management Science from Northwestern University. Her research interests include simulation input and output uncertainty analysis, meta-modeling, and simulation-based optimization. Her email address is linyujin@amazon.com.

JINGTAO ZHANG is a Research Scientist from Supply Chain Optimization Technology team in Amazon. He received his Ph.D. degree in Industrial and Systems Engineering from Virginia Tech. His research interests include design and analysis of stochastic simulation experiments, global sensitivity analysis, and simulation optimization. His email address is jingtaz@amazon.com.

MITCHELL PERRY is a Research Scientist from the Supply Chain Optimization Technology team in Amazon. He received his Ph.D. degree in Operations Research from Columbia University. His research interests include applied probability and optimization. His email address is perrymit@amazon.com.

XIAOYU LU is an Applied Scientist from Supply Chain Optimization Technology team in Amazon. She received her Ph.D. degree in Statistical Science from University of Oxford. Her research interests include machine learning, Bayesian inference, reinforcement learning and generative models. Her email address is luxiaoyu@amazon.com.

YUNAN LIU is a Principal Research Scientist from Supply Chain Optimization Technology team in Amazon. He earned his Ph.D. in Operations Research from Columbia University. His research interests include stochastic modeling, simulation, optimal control and reinforcement learning, with applications to queueing and supply chain systems. His email address is yunanliu@amazon.com. His website is https://yliu48.github.io/.

HOIYI NG is a Principal Research Scientist from Supply Chain Optimization Technology team in Amazon. She received her MS degree in Statistics from University of Washington, Seattle. Her research interests include causal inference, graphical causal models, and the intersection between causal inference and large language models. Her email address is nghoiyi@amazon.com.