Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

DIGITAL TWIN VALIDATION WITH MULTI-EPOCH, MULTI-VARIATE OUTPUT DATA

Linyun He[1], Luke Rhodes-Leader[2], and Eunhye Song[1]

[1]School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[2]Dept. of Management Science, Lancaster University, Lancaster, UK

ABSTRACT

This paper studies validation of a simulation-based process digital twin (DT). We assume that at any point the DT is queried, the system state is recorded. Then, the DT simulator is initialized to match the system state and the simulations are run to predict the key performance indicators (KPIs) at the end of each time epoch of interest. Our validation question is if the distribution of the simulated KPIs matches that of the system KPIs at every epoch. Typically, these KPIs are multi-variate random vectors and non-identically distributed across epochs making it difficult to apply the existing validation methods. We devise a hypothesis test that compares the marginal and joint distributions of the KPI vectors, separately, by transforming the multi-epoch data to identically distributed observations. We empirically demonstrate that the test has good power when the system and the simulator sufficiently differ in distribution.

1 INTRODUCTION

Process Digital Twins (DTs) continue to grow in popularity as a decision-making tool in industries, while stochastic simulation models are increasingly seen as a potential base model for DT in many applications (Biller et al. 2022). The key difference between simulation DT and a traditional simulator is that the DT is continuously updated to keep up with the changes in the target system to enable context-specific predictions and decisions. Here, a context is the state of the system or the decision to be implemented at some time point when the DT is queried to predict the system performance in the near future; we refer to this time window as an epoch in this paper. The context determines the initial conditions of the DT simulator so that the simulated key performance indicators (KPIs) at the end of the epoch provides valid prediction for the system KPIs provided that the simulation model approximates the system well.

Since misalignment between the system and the DT may impact future decisions and the system performance, it is important to quickly detect it if the simulated KPIs start deviating significantly from the KPIs observed from the system. However, the context-dependency of the DT use brings significant challenges to validation of the DT simulator (Lugaresi et al. 2023). That is, the differences in initial conditions of the epochs makes it difficult to apply traditional validation procedures that assume that the underlying distribution of the system outputs remains unchanged across epochs and multiple observations at each epoch are available. Moreover, in a realistic system, there are several statistically dependent KPIs. Simply validating each KPI marginally ignores the potentially important dependence among the KPIs. For example, in a manufacturing system, understanding the dependence between the delays at two stations is crucial for efficiently managing the workflow. *This paper proposes a validation method that utilises multi-variate KPIs observed from the system and simulated by the DT from multiple epochs.*

There is a stream of literature on simulation model validation that relies on hypothesis tests to determine whether the KPIs observed from the system and generated by the simulator are identically distributed (Naylor and Finger 1967; Shannon 1976; Schruben 1980; Balci and Sargent 1981). These methods require a large number of independent and identically distributed (i.i.d.) system observations (Sargent 2015). On the other hand, in the DT setting, system-observed KPIs at multiple epochs are rarely i.i.d. which makes it difficult to extend the existing methods to pool multi-epoch data for model validation.

More recent approaches compare the real and simulated time series. Hua et al. (2022) apply trace-driven simulation (Kleijnen et al. 1998) to enable the validation, however, this only tests the deterministic model logic, not the stochastic input models of the simulator. Addressing the shortcoming, Lugaresi et al. (2019) propose the quasi Trace-Driven Simulation (qTDS), where observed random quantities are modified through the simulation input models. They apply spectral density estimation to the two time-series observed from the system and generated by qTDS to measure their difference. Adopting qTDS, Lugaresi et al. (2023) propose subsequence measures for the comparison such as the longest common subsequence and dynamic time warping, which produce a statistic that can be compared against a threshold. However, the threshold can be difficult to set and interpret. The experiments on a small manufacturing line suggest the methods are able to detect changes in mean processing times at bottlenecks, but are less sensitive to changes in variance or differences not at a bottleneck. Morgan and Barton (2022) take a discrete Fourier transform of both time series and propose a hypothesis test on the Fourier coefficients. The use of a statistical test allows a threshold to be selected based on the required confidence level. Each of these methods consider a single epoch at a time, with the focus on whether the simulation model is aligned within the epoch.

When multi-epoch data is available, it is natural to utilise them to achieve a greater degree of statistical confidence. However, the non-i.i.d. nature of the data imposes challenges. Oakley et al. (2020) propose a method that assumes that conditional on the initial states, the relative system trajectories across multiple epochs are i.i.d. This approximation may hold for some systems, but may not make sense in the DT applications.

All aforementioned methods assume that there is only one KPI of interest, not multi-variate. dos Santos et al. (2023) propose a method that can handle multi-variate data by collecting an equal number of observations, n, from the system and the simulator and then train a k-nearest-neighbour classifier on the results. The accuracy of the classifier is then monitored for each set of n observations using a control chart. However, their approach requires more granular system data than those discussed in this paper. For instance, their experiment includes a factory DT in which the processing and waiting times of all parts are the observations from a single epoch, where the control chart's thresholds are set from the past epoch's data. In this research, we assume that only system-level KPIs are measured, not part-level.

We propose a hypothesis test that aggregates multi-epoch, multi-variate KPIs to test if the distributions of the simulated and system KPIs match in all epochs. The test is decomposed to a set of marginal tests and a joint test. The former examines if each marginal KPI's distributions match applying the hypothesis testing framework proposed by Rhodes-Leader and Nelson (2023) designed for one-dimensional KPIs. For the joint test, we impose an assumption that the copula of the marginal distributions in all epochs are identical. Under this assumption, we first transform each system KPI at each epoch using the marginal cumulative distribution function (cdf) of the simulated KPI. Then, the resulting transformed random vector has the same copula in all epochs under the null. This allows us to compute a test statistic by aggregating all epochs' data. Lastly, we adopt the multiple hypothesis testing framework proposed by Romano and Wolf (2005) to lessen the conservatism from achieving the family-wise error rate (FWER).

The paper is structured as follows. Section 2 defines the scope of our DT simulator calibration problem. The test for the marginals is reviewed in Section 3 and the proposed copula-based test is described in Section 4. Section 5 discusses how to combine the marginal and joint tests into the multiple hypothesis testing framework. Empirical results from a controlled experiment are presented in Section 6.

2 PROBLEM DEFINITION

To describe the multi-epoch data collected from the system and the DT simulator, we first establish notation with a system illustration in Figure 1. Let subscript i label the ith epoch. We denote the start time and duration of the epoch by t_i and Δ_i, respectively. At the beginning of each epoch, the system records the snapshot of the system state (e.g., work-in-progress—WIP—level, elapsed processing times, operator/tool status at each station, etc.), which can be summarized by the state vector, Ψ_i. The d-dimensional system KPI, $\mathbf{W}_i = (W_{i1}, \ldots, W_{id})$, are collected at the end of the epoch, i.e., at time $t_i + \Delta_i$. We also denote the

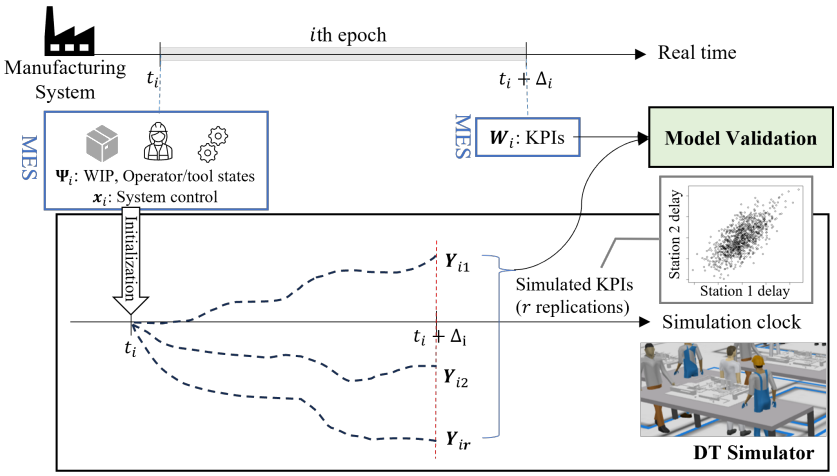

Figure 1: An illustration of the system and simulated KPIs at the ith epoch.

system control adopted during the ith epoch by $\mathbf{x}_i$, and the collection of random realizations (e.g., processing times) occurred during the epoch by ξ_i. Then, $\mathbf{W}_i$ is a function of $\Psi_i, \mathbf{x}_i$, and ξ_i.

Typically, Ψ_i and ξ_i are random in the target system, and $\mathbf{x}_i$ is chosen based on Ψ_i. Also, only one realization of (Ψ_i, ξ_i) is observed for each i. However, considering the randomness in ξ_i, we can view $\mathbf{W}_i = \mathbf{W}_i(\Psi_i, \mathbf{x}_i, \xi_i)$ as a random vector with cumulative distribution function (cdf) $F_i(\cdot|\Psi_i, \mathbf{x}_i)$. Notice that the distribution is conditional on Ψ_i and $\mathbf{x}_i$ clearly indicating the within-epoch KPI's dependence on the initial state of the epoch. Therefore, $\mathbf{W}_i$ is not identically distributed across all is in general. To address this, we introduce the following assumption, where n denotes the total number of epochs and $[n] = \{1, 2, \ldots, n\}$.

Assumption 1 $\{\mathbf{W}_i\}_{i\in[n]}$ are independent conditional on the initial states $\{\Psi_i\}_{i\in[n]}$ and controls $\{\mathbf{x}_i\}_{i\in[n]}$.

For each epoch, we refer to the initial state and control together as the epoch's context. Assumption 1 posits that the dependence among $\{\mathbf{W}_i\}_{i\in[n]}$ arises from their dependence on the contexts. Hereafter, all probability statements we make are conditional on the observed contexts.

We assume that the DT simulator is set up so that the KPI observed at the end of each epoch i can be emulated by initializing the simulator's state variables to Ψ_i and adopting the same control, $\mathbf{x}_i$, as in the system. What corresponds to ξ_i observed from the system is the vector of random inputs, ω_i, generated by the DT simulator. Similarly, the simulated KPI at epoch i, $\mathbf{Y}_i$, is a function of the initial state, the control, and within-epoch random realizations, i.e., $\mathbf{Y}_i = \mathbf{Y}_i(\Psi_i, \mathbf{x}_i, \omega_i)$, and we denote its distribution conditional on Ψ_i and $\mathbf{x}_i$ by $G_i(\cdot|\Psi_i, \mathbf{x}_i)$, where the randomness comes from ω_i. As in Assumption 1, we assume that the simulation outputs are independent across the epochs given the contexts. Furthermore, we assume that both $\mathbf{W}_i$ and $\mathbf{Y}_i$ are continuous random vectors in all epochs.

Since no model is perfect, the distributions of $\mathbf{W}_i$ and $\mathbf{Y}_i$ conditional on Ψ_i and $\mathbf{x}_i$ are unlikely to be equal in general. The discrepancy may be due to logical error, mismatch in the distributions of ξ_i and ω_i, and/or some parameter settings of the simulation model. Nevertheless, the hope is that the DT can closely mimic the target systems' KPIs in *all epochs* so that predictive and prescriptive analyses derived from the DT outputs will be meaningful. To this end, the first step is to validate the simulation model. At the high level, we aim to test the following hypothesis:

$$H_0 : F_i(\cdot|\Psi_i, \mathbf{x}_i) = G_i(\cdot|\Psi_i, \mathbf{x}_i) \text{ for all } i \in [n]. \tag{1}$$

There are two challenges at testing (1). The first is that it involves multi-epoch KPIs collected from the system and the simulator that are not identically distributed across all epochs. In particular, we have only one observation of the system KPI vector, $\mathbf{W}_i$, from each epoch i, which makes it difficult to estimate

F_i. Secondly, we are interested in multi-dimensional KPI in this work, which further complicates the test as the joint distribution of d elementwise KPIs must be considered.

The approach we take in this paper is to construct a multiple hypothesis testing that examines if each set of marginal distributions (system and simulated KPI's) as well as the joint distributions match by converting the non-identically distributed data to be approximately identically distributed via transformation.

3 HYPOTHESIS TESTING FOR MARGINAL PERFORMANCE MEASURE

As part of the validation process, we propose that each dimension of the KPI is tested marginally applying the framework proposed by Rhodes-Leader and Nelson (2023).

Unlike the target system, one can run $r > 1$ replications of the simulation at each epoch. These simulation outputs can be then used to compute statistical risk measures such as the probability of completing a job within a certain time window, or the percentile of the number of jobs waiting at each station. Let $\{\mathbf{Y}_{ij}\}_{j\in[r]}$ represent the collection of r simulated KPIs from the ith epoch, where $\mathbf{Y}_{ij} = (Y_{ij1}, \ldots, Y_{ijd})$.

Consider the lth performance measure. At the end of the nth epoch, we have $\{W_{il}\}_{i\in[n]}$ from the real system and r replications of simulated performance measure in each epoch, $\{Y_{ijl}\}_{i\in[n],j\in[r]}$. Let $F_{il}(\cdot|\Psi_i, \mathbf{x}_i)$ and $G_{il}(\cdot|\Psi_i, \mathbf{x}_i)$ be the lth marginal distribution of $\mathbf{W}_i$ and $\mathbf{Y}_{ij}$, respectively. We conduct hypothesis testing to determine whether the marginal distributions are identical across all n epochs. Namely, the null hypothesis for the lth test is stated as follows

$$H_0^l : F_{il}(\cdot|\Psi_i, \mathbf{x}_i) = G_{il}(\cdot|\Psi_i, \mathbf{x}_i) \text{ for all } i \in [n]. \tag{2}$$

From the probability integral transform, $G_{il}(W_{il}|\Psi_i, \mathbf{x}_i)$ is uniformly distributed under the null hypothesis. Thus, $\{G_{il}(W_{il}|\Psi_i, \mathbf{x}_i)\}_{i\in[n]}$ is an i.i.d. sample from $\mathrm{U}(0,1)$ under H_0^l, which can be tested for uniformity using a standard goodness-of-fit test, such as Kolmogorov–Smirnov (KS) and Anderson-Darling (AD) tests.

Let $U_{(1)} \le U_{(2)} \le \cdots \le U_{(n)}$ be the order statistics of $\{G_{il}(W_{il}|\Psi_i, \mathbf{x}_i)\}_{i\in[n]}$. Then, the KS test rejects the null hypothesis of uniformity if the statistic, $D_{KS} = \max_{i\in[n]}(\frac{i}{n} - U_{(i)}, U_{(i)} - \frac{i-1}{n})$, exceeds a critical value. Similarly, the AD test computes $D_{AD} = -n - \sum_{i=1}^{n} \frac{2i-1}{n}(\ln(U_{(i)}) + \ln(1 - U_{(n+1-i)}))$ and rejects the null hypothesis if the statistic is greater than a critical value.

Although $G_{il}(\cdot|\Psi_i, \mathbf{x}_i)$ is unknown, it can be estimated to an arbitrary precision using a sufficiently large number of replications, r, of the DT simulator and setting up the empirical cdf (ecdf) of the observations of the lth KPI for the ith epoch:

$$\hat{G}_{il}(y) = \frac{1}{r}\sum_{j=1}^{r} \mathbf{1}\left\{Y_{ijl} \le y\right\}, \tag{3}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. For sufficiently large r, $\hat{G}_{il}(W_{il})$ is approximately uniformly distributed under H_0^l. In the empirical study in Section 6, we modify (3) to prevent infinite values when calculating the logarithmic terms in D_{AD}.

Since there are $d > 1$ KPIs of interest, the null hypothesis for matching all marginal distributions is:

$$H_0^M : H_0^l \text{ holds for all } l \in [d].$$

To test H_0^M, the FWER must be carefully controlled. In general, the d KPIs are statistically dependent with an unknown dependence structure. A standard Bonferroni correction can be adopted, however, is known to be conservative, which diminishes the power of the test, i.e., the probability of correctly rejecting the null hypothesis when differences do exist is lower. This reduces the ability of the test to detect a misalignment between the real system and the simulator. Section 5 discusses a multiple hypothesis testing framework that reduces such conservatism.

4 COPULA HYPOTHESIS TESTING FOR JOINT DISTRIBUTION

In this section, we discuss a hypothesis test on the joint distributions of $\mathbf{W}$ and $\mathbf{Y}$ assuming their element-wise marginal distributions match. Under this assumption, we characterize their respective joint distributions by copulas and utilize the copula goodness-of-fit test for validation.

A d-dimensional copula is a joint distribution of d random variables marginally distributed as $U(0,1)$. Sklar's theorem states that for any d-dimensional random vector $\mathbf{X}$ with cdf $H : \mathbb{R}^d \to [0,1]$, there exists copula C that satisfies $H(x_1,\ldots,x_d) = C(H_1(x_1),\ldots,H_d(x_d))$ for all $(x_1,\ldots,x_d) \in \mathbb{R}^d$, where H_ℓ is the marginal cdf of the ℓth element of $\mathbf{X}$ (Durante et al. 2013). Moreover, for continuous $\mathbf{X}$, the copula, C, is uniquely defined by H and the marginal distributions. Conversely, given the d marginals and the copula, the joint distribution is uniquely determined. Thanks to these properties, the copula has been adopted as a powerful tool for characterizing a joint distribution of random variables.

Let $U_l = F_{il}(W_{il}|\Psi_i, \mathbf{x}_i)$ for $l \in [d]$. Then, the distribution of $(U_1,\ldots,U_d)$ defines unique copula $C_i^{\mathbf{W}}(\mathbf{u}) \triangleq \mathbb{P}(U_1 \le u_1,\ldots,U_d \le u_d)$, where $\mathbf{u} = (u_1,\ldots,u_d) \in [0,1]^d$. We define the copula, $C_i^{\mathbf{Y}}$, from the marginal and joint distributions of $\mathbf{Y}_{ij}$ in a similar fashion. Suppose the marginal distributions of $\mathbf{W}_i$ and $\mathbf{Y}_i$ match in all epochs. Then, we can adopt the following null hypothesis to compare their joint distributions:

$$H_0^C : C_i^{\mathbf{W}} = C_i^{\mathbf{Y}} \text{ for all } i \in [n]. \tag{4}$$

Namely, H_0^C implies that the copulas constructed form $\mathbf{W}_i$ and $\mathbf{Y}_i$ match in each epoch.

Clearly, the analytical expressions for $C_i^{\mathbf{W}}$ and $C_i^{\mathbf{Y}}$ are unknown in general. Since the r simulation replications $\{\mathbf{Y}_{ij}\}_{j\in[r]}$ are i.i.d. within each epoch, we can construct an estimator for $C_i^{\mathbf{Y}}$ for each $i \in [n]$. On the other hand, we have only one observation of $\mathbf{W}_i$ from the ith epoch, which makes it challenging to construct a test statistic for H_0^C. To facilitate the hypothesis testing, we make the following assumption.

Assumption 2 Conditional on $\{\Psi_i\}_{i\in[n]}$ and $\{\mathbf{x}_i\}_{i\in[n]}$, $C_i^{\mathbf{W}}$ and $C_i^{\mathbf{Y}}$ remain unchanged for all $i \in [n]$.

Assumption 2 is clearly weaker than assuming F_i is identical for all $i \in [n]$; the marginal distributions of the KPIs may still vary by the epoch. Moreover, it allows $\{W_i\}_{i\in[n]}$ and $\{Y_{ij}\}_{i\in[n],j\in[r]}$ to be dependent as long as Assumption 1 holds. Under Assumption 2, we can now pool $\{\mathbf{W}_i\}_{i\in[n]}$ to estimate $C_i^{\mathbf{W}}$.

To construct the test statistics for H_0^C, we start by introducing the estimator for $C_i^{\mathbf{Y}}$. For each epoch i, by transforming $\{\mathbf{Y}_{ij}\}_{j\in[r]}$ using the marginal ecdfs given in (3), we define $\hat{\mathbf{V}}_{ij} = (\hat{V}_{ij1},\ldots,\hat{V}_{ijd}) \triangleq (\hat{G}_{i1}(Y_{ij1}),\ldots,\hat{G}_{id}(Y_{ijd}))$ and $\mathbf{Z}_{ij}^{\mathbf{Y}} = (z_{ij1}^{\mathbf{Y}},\ldots,z_{ijd}^{\mathbf{Y}}) \triangleq \frac{r}{r+1}\hat{\mathbf{V}}_{ij} = \frac{r}{r+1}(\hat{G}_{i1}(Y_{ij1}),\ldots,\hat{G}_{id}(Y_{ijd}))$ for $j \in [r]$. We refer to $\{\mathbf{Z}_{ij}^{\mathbf{Y}}, j \in [r]\}$ as the pseudo-sample from the ith epoch generated by $\mathbf{Y}_i$, while each $\mathbf{Z}_{ij}^{\mathbf{Y}}$ as a pseudo-vector. Note that $r\hat{G}_{il}(Y_{ijl})$ is the marginal rank of the lth simulated KPI observed from the jth replication among all r replications. The scaling term, $r/(r+1)$, ensures that $\mathbf{Z}_{ij}^{\mathbf{Y}}$ is contained in the interior of $[0,1]^d$. Since simulated KPIs from all r replications are utilized to estimate $\hat{G}_{il}$ for each l, $\{\mathbf{Z}_{ij}^{\mathbf{Y}}\}_{j\in[r]}$ are no longer i.i.d. Nevertheless, when r is sufficiently large, $\hat{G}_{il}$ closely approximates G_{il} and the dependence among $\mathbf{Z}_{ij}^{\mathbf{Y}}, j \in [r]$ becomes weaker. Consequently, the psuedo sample behaves more closely to an i.i.d. observation from $C_i^{\mathbf{Y}}$. From this intuition, Deheuvels (1979) introduces the empirical copula estimator:

$$\hat{C}_i^{\mathbf{Y}}(\mathbf{u}) \triangleq \frac{1}{r+1}\sum_{j=1}^{r} \mathbf{1}\{z_{ij1}^{\mathbf{Y}} \le u_1,\ldots,z_{ijd}^{\mathbf{Y}} \le u_d\} \text{ for } i \in [n]. \tag{5}$$

Under Assumption 2, $\{\hat{\mathbf{V}}_{ij}\}_{i\in[n],j\in[r]}$ are identically distributed, therefore, we can further improve (5) by pooling all $\hat{\mathbf{V}}_{ij}$'s. Let $\{\mathbf{Z}_{ij}^{\mathbf{Y}}\}_{i\in[n],j\in[r]}$ be the new pseudo-sample after pooling, where for each $l \in [d]$, $z_{ijl}^{\mathbf{Y}} = \frac{R_{ijl}}{nr+1}$ and R_{ijl} is redefined as the rank of $\hat{V}_{ijl}$ among $\{\hat{V}_{ikl}\}_{i\in[n],k\in[r]}$. The resulting estimator of $C^{\mathbf{Y}}$ is

$$\hat{C}^{\mathbf{Y}}(\mathbf{u}) \triangleq \frac{1}{nr+1}\sum_{i=1}^{n}\sum_{j=1}^{r} \mathbf{1}\{z_{ij1}^{\mathbf{Y}} \le u_1,\ldots,z_{ijd}^{\mathbf{Y}} \le u_d\}.$$

Next, we discuss estimation of $C_i^{\mathbf{W}}$. Since we only observe one $\mathbf{W}_i$ for each i, we do not have estimators for $F_{il}, l \in [d]$. However, under H_0^M, we can estimate F_{il} with $\hat{G}_{il}$. From this, let $\hat{\mathbf{U}}_i = (\hat{U}_{i1}, \ldots, \hat{U}_{id}) = (\hat{G}_{i1}(W_{i1}), \ldots, \hat{G}_{id}(W_{id}))$. From Assumption 2, $\{\hat{\mathbf{U}}_i\}_{i\in[n]}$ are i.i.d, therefore, we can compute the pseudo-samples $\{\mathbf{Z}_i^{\mathbf{W}}\}_{i\in[n]}$ out of $\{\hat{\mathbf{U}}_i\}_{i\in[n]}$ by taking $z_{il}^{\mathbf{W}} = \frac{1}{n+1}\sum_{k=1}^{n} \mathbf{1}\{\hat{U}_{kl} \le \hat{U}_{il}\}$. Then, we can estimate $\hat{C}^{\mathbf{W}}$ in a similar way as in (5):

$$\hat{C}^{\mathbf{W}}(\mathbf{u}) \triangleq \frac{1}{n+1}\sum_{i=1}^{n} \mathbf{1}\{z_{i1}^{\mathbf{W}} \le u_1, \ldots, z_{id}^{\mathbf{W}} \le u_d\}.$$

We note that in the classical copula goodness-of-fit literature, the pseudo-samples $\{\mathbf{Z}_i^{\mathbf{W}}\}_{i\in[n]}$ are computed from ranks of the raw data $\{\mathbf{W}_i\}_{i\in[n]}$ rather than from those of the transformed vectors $\{\hat{\mathbf{U}}_i\}_{i\in[n]}$. We make this modification because the pseudo-samples should ideally be computed from i.i.d. data. However, in our case, the $\mathbf{W}_i$'s may not be identically distributed, while the transformed $\hat{\mathbf{U}}_i$'s are, under Assumption 2 and H_0^M. For a similar reason, we favor estimating $\hat{C}^{\mathbf{Y}}$ from the transformed $\hat{\mathbf{V}}_{ij}$'s rather than $\mathbf{Y}_{ij}$'s. By pooling the $\hat{\mathbf{V}}_{ij}$'s across epochs to compute the pseudo-vectors, we enlarge the sample size used in the copula estimation. This enhancement improves the estimation convergence rate, albeit at the cost of compromising the independence among the $\hat{\mathbf{V}}_{ij}$'s.

To measure the discrepancy between the two estimated copulas, we select three test statistics from the literature, construct the copula hypothesis tests based on them, which are empirically compared in Section 6. For a comprehensive review of the copula goodness-of-fit testing, see Berg (2009).

Genest and Rémillard (2008) propose a Cramér–von Mises (CvM)-type statistics based on the pseudo-vector copulas that measures

$$\hat{T}_1 = n\int_{[0,1]^d} \left(\hat{C}^{\mathbf{W}}(\mathbf{u}) - \hat{C}^{\mathbf{Y}}(\mathbf{u})\right)^2 \mathrm{d}\hat{C}^{\mathbf{W}}(\mathbf{u}) = \sum_{i=1}^{n} \left(\hat{C}^{\mathbf{W}}(\mathbf{Z}_i^{\mathbf{W}}) - \hat{C}^{\mathbf{Y}}(\mathbf{Z}_i^{\mathbf{W}})\right)^2,$$

which is the expected squared difference between $\hat{C}^{\mathbf{W}}(\mathbf{u})$ and $\hat{C}^{\mathbf{Y}}(\mathbf{u})$ with respect to $\hat{C}^{\mathbf{W}}(\mathbf{u})$.

The second approach involves projecting the d-dimensional copulas to one dimension using Kendall's dependence function, $K(u) = \mathbb{P}(C(\mathbf{Z}) \le u), u \in [0,1]$. The sample versions for $\mathbf{W}$ and $\mathbf{Y}$ are $\hat{K}^{\mathbf{W}}(u) = \frac{1}{n+1}\sum_{i=1}^{n} \mathbf{1}\{\hat{C}^{\mathbf{W}}(\mathbf{Z}_i^{\mathbf{W}}) \le u\}$ and $\hat{K}^{\mathbf{Y}}(u) = \frac{1}{nr+1}\sum_{i=1}^{n}\sum_{j=1}^{r} \mathbf{1}\{\hat{C}^{\mathbf{Y}}(\mathbf{Z}_{ij}^{\mathbf{Y}}) \le u\}$, respectively. Genest et al. (2006) propose to compute the CvM-type test statistic on the distance between $K^{\mathbf{W}}$ and $K^{\mathbf{Y}}$:

$$\hat{T}_2 = n\int_0^1 \left(\hat{K}^{\mathbf{W}}(u) - \hat{K}^{\mathbf{Y}}(u)\right)^2 \mathrm{d}\hat{K}^{\mathbf{W}}(u).$$

Unlike the first two test statistics that measure the distance between the empirical copulas, Panchenko (2005) propose to measure the distance between the pseudo-samples using a positive definite kernel, $k_d(\cdot,\cdot)$. In our experiments, we adopt the Gaussian kernel: $k_d(\mathbf{S}_1, \mathbf{S}_2) \triangleq \exp(-\|\mathbf{S}_1 - \mathbf{S}_2\|^2/(2dh^2))$, where $\mathbf{S}_1$ and $\mathbf{S}_2$ are d-dimensional vectors, $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^d$ and h is a bandwidth parameter. For two d-dimensional integrable functions h_1 and h_2, an inner product of h_1 and h_2 can be defined from the Gaussian kernel as $\langle h_1|k_d|h_2\rangle \triangleq \int\int k_d(\mathbf{s}_1,\mathbf{s}_2)h_1(\mathbf{s}_1)h_2(\mathbf{s}_2)\mathrm{d}\mathbf{s}_1\mathrm{d}\mathbf{s}_2$. The inner product induces a squared distance Q: $Q(h_1,h_2) \triangleq \langle h_1 - h_2|k_d|h_1 - h_2\rangle = \langle h_1|k_d|h_1\rangle + \langle h_2|k_d|h_2\rangle - 2\langle h_1|k_d|h_2\rangle$. If h_1 and h_2 are probability densities of $\mathbf{S}_1$ and $\mathbf{S}_2$, respectively, then the inner product $\langle h_1|k_d|h_2\rangle$ is identical to the expectation, $\mathbb{E}[k_d(\mathbf{S}_1,\mathbf{S}_2)]$. Consequently, we have $Q(h_1,h_2) = \mathbb{E}[k_d(\mathbf{S}_1,\mathbf{S}_1)] + \mathbb{E}[k_d(\mathbf{S}_2,\mathbf{S}_2)] - 2\mathbb{E}[k_d(\mathbf{S}_1,\mathbf{S}_2)]$.

Applying Q to measure the distance between the probability densities of $\mathbf{Z}_i^{\mathbf{W}}$ and $\mathbf{Z}_{ij}^{\mathbf{Y}}$, we have the third test statistic

$$\hat{T}_3 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\ell=1}^{n} k_d(\mathbf{Z}_i^{\mathbf{W}}, \mathbf{Z}_\ell^{\mathbf{W}}) + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{\ell=1}^{n} k_d(\tilde{\mathbf{Z}}_i^{\mathbf{Y}}, \tilde{\mathbf{Z}}_\ell^{\mathbf{Y}}) - \frac{2}{n^2}\sum_{i=1}^{n}\sum_{\ell=1}^{n} k_d(\mathbf{Z}_i^{\mathbf{W}}, \tilde{\mathbf{Z}}_\ell^{\mathbf{Y}}), \tag{6}$$

where $\{\tilde{\mathbf{Z}}_i\}_{i\in[n]}$ is a size-n random sample uniformly drawn from the pseudo-vectors, $\{\mathbf{Z}_{ij}^{\mathbf{Y}}\}_{i\in[n],j\in[r]}$. Notice that the expectations in the definition of Q are replaced with the sample averages in (6). As we have nr pseudo-vectors of the simulated KPIs and only n system KPIs, we subsample the former in computing $\hat{T}_3$.

To perform the hypotheses tests using the three test statistics defined above, we first estimate their critical values via bootstrapping at a given confidence level α. Let B be the bootstrap size. In the ith epoch, we bootstrap B sets of $r+1$ simulation outputs, $\{\mathbf{Y}_{ij}^{*b}\}_{j\in[r+1]}$ for $b \in [B]$, by sampling $\{\mathbf{Y}_{ij}\}_{j\in[r]}$ with replacement. We then implement the test procedure independently on each $\{\mathbf{Y}_{ij}^{*b}\}_{j\in[r+1]}$ data set. Namely, for each $b \in [B]$, we compute the marginal ecdfs $\hat{G}_{il}^b$ from replications $\{\mathbf{Y}_{ij}^{*b}\}_{j\in[r]}$ and compute $\hat{\mathbf{V}}_{ij}^b = (\hat{V}_{ij1}^b, \ldots, \hat{V}_{ijd}^b) = (\hat{G}_{i1}^b(Y_{ij1}^{*b}), \ldots, \hat{G}_{id}^b(Y_{ijd}^{*b}))$ for $j \in [r]$. We take $\mathbf{Y}_{i(r+1)}^{*b}$ in the place of $\mathbf{W}_i$ and compute $\hat{\mathbf{U}}_i^b = (\hat{U}_{i1}^b, \ldots, \hat{U}_{id}^b) = (\hat{G}_{i1}(Y_{i(r+1)1}^{*b}), \ldots, \hat{G}_{id}^b(Y_{i(r+1)d}^{*b})$. We proceed by calculating the pseudo-samples and corresponding statistics $\hat{T}^b$ for each of the three tests. Finally, we compute the $1-\alpha$ empirical quantile among $\{\hat{T}^b\}_{b\in[B]}$ as the critical value for the test. Let $\hat{T}$ be the statistic calculated from the original $\{\mathbf{W}_i\}_{i\in[n]}$ and $\{\mathbf{Y}_{ij}\}_{i\in[n],j\in[r]}$. The null hypothesis (4) is rejected if $\hat{T}$ exceeds the quantile.

5 MULTIPLE HYPOTHESIS TESTING

To summarize our discussion thus far, H_0 in (1) consists of $d+1$ hypotheses, H_0^l for $l \in [d]$ and H_0^C. If each of these are tested at α significance, the probability that at least one of the true hypotheses is rejected (FWER) is greater than or equal to α.

To achieve α-level FWER at α for testing H_0 in (1), we first split the error rate between the marginal and joint tests as $\alpha_M = \frac{d}{d+1}\alpha$ and $\alpha_J = \alpha/(d+1)$, respectively. Then, the Bonferroni correction guarantees that the FWER is bounded away from α. However, the Bonferroni correction can be very conservative. Applying it to all d marginal tests makes each test have $\alpha/(d+1)$ significance level. This implies that the null is much more likely to be accepted, and consequently reduces the full procedure's ability to detect misalignment.

To lessen the conservatism, we adopt a stepdown procedure for multiple hypothesis tests proposed by Romano and Wolf (2005) for the marginal tests. It is a sequential procedure that tests a series of joint hypotheses, each time potentially rejecting a hypothesis and removing it from consideration. The aim of this procedure is to exert strong control over the FWER whilst accounting for dependence between tests.

The first step is to calculate the test statistic for each of the marginal goodness-of-fit tests, $\{T_l\}_{l\in[d]}$, and sort them in decreasing order: $T_{(1)} \geq T_{(2)} \geq \ldots \geq T_{(d)}$. For simplicity, assume their corresponding labels are $l = 1, \ldots, d$. Let $c_d(\alpha)$ denote the critical value of the maximum of d statistics at α confidence. We then test $T_{(1)}$ against the critical value $c_d(1-\alpha_M)$. If $T_{(1)} \leq c_d(1-\alpha_M)$, then the procedure does not reject any of the null hypotheses and we conclude there is no statistical evidence to reject H_0^M. If $T_{(1)} > c_d(1-\alpha_M)$, then we reject H_0^1, redefine H_0^M to only include the remaining $d-1$ hypotheses, $H_0^M = \{H_0^l : 2 \leq l \leq d\}$, and move on to testing H_0^M. That is, we test $T_{(2)}$, the maximum of $d-1$ statistics, against the corresponding critical value, $c_{d-1}(1-\alpha_M)$. As before, if $T_{(2)} \leq c_{d-1}(1-\alpha_M)$, do not reject the (new) null hypothesis, H_0^M, then the procedure stops. Otherwise, it rejects H_0^2 and continues in the same manner. In Theorem 1 of Romano and Wolf (2005), this stepdown procedure is proved to have strong control of the FWER, under the assumption that $c_l(1-\alpha_M)$ increases with l.

The critical values can be estimated using resampling techniques such as bootstrapping (Romano and Wolf 2005). For each bootstrap sample, we first compute a full set of $\{T_l\}_{l\in[d]}$. To estimate $c_d(1-\alpha_M)$, for instance, $T_{(1)}$ is calculated from each sample and the critical value is estimated by the $1-\alpha_M$ empirical quantile of the $T_{(1)}$ values. Similarly, $c_l(1-\alpha_M)$ is estimated by the $1-\alpha_M$ empirical quantile of the $T_{(d+1-l)}$ values from the same bootstrap samples. This procedure avoids independent resampling in each step and is computationally efficient; it also accounts for the dependence among the $\{T_l\}_{l\in[d]}$ statistics.

One can include the joint test within the stepdown procedure. However, as Romano and Wolf (2005) remark, their stepdown procedure works best when each of the tests is balanced (from the power perspective). The scales of the marginal and joint test statistics are not necessarily the same, which may lead to unbalanced tests. There are methods to achieve the balance, such as studentization, but these would require an additional layer of bootstrapping. Thus, for simplicity, we do not apply this routine.

Although a single rejection from the marginal tests is sufficient to reject H_0, and so we could stop at this point having declared the simulation model invalid, we continue through all of the marginal tests. Identifying which of the KPIs are not being adequately predicted can be instructive when it comes to model correction and calibration. We do not proceed to the copula test in this circumstance, as the validity of the joint test is dependent on the validity of the marginal distributions.

6 EMPIRICAL STUDY

In this section, we examine the performance of the multiple hypothesis testing framework we propose using a toy example set up with multivariate Gaussian random vectors.

Suppose that in the ith epoch, the lth system KPI, W_{il}, follows a normal distribution: $W_{il} \sim \mathcal{N}(A_{il} + \varepsilon\sqrt{l}, \delta l B_{il}^2)$. Here, A_{il} and B_{il} are components of the initial state vector Ψ_i, with $A_{il} \sim \mathcal{N}(2\sqrt{l}, 0.5^2)$ and $B_{il} \sim \mathcal{N}(2, 0.5^2)$. For the DT simulator, we set $Y_{il}|\Psi_i \sim \mathcal{N}(A_{il}, lB_{il}^2)$. Therefore, ε and δ, characterize the discrepancy between the system KPIs and the simulated KPIs in their marginals. In our study, we adjust the values of ε and δ to control the extent of discrepancy.

Let $D(\theta)$ be the correlation matrix of $\mathbf{W}_i$, with $[D(\theta)]_{kl}$ denoting its (k,l)th entry. By definition, each diagonal element $[D(\theta)]_{ll} = 1$ for $l \in [d]$. We make each off-diagonal entry identical by setting $[D(\theta)]_{kl} = \theta\frac{d-1}{d}$ for $k \neq l$. Here, θ is a constant that controls the level of discrepancy in the joint distributions of the KPI vectors, which subsequently affects the discrepancy of copulas. We assume that $D(\theta)$ remains unchanged for $\mathbf{W}_i$ across all epochs, i.e., Assumption 2 holds in this example. For the DT simulator, we set the correlation matrix of $\mathbf{Y}_{ij}$ as $[D(\frac{1}{2})]$ for all $i \in [n]$ and $j \in [r]$. That is, when $\theta = 1/2$, the correlation matrices of the system and the DT simulator match. The farther θ moves away from $1/2$, the more discrepancy in the joint distribution is induced.

To study the power of our test under various degrees of discrepancy, we choose ε, δ and θ at different levels. Since ε changes the mean vector of $\mathbf{W}_i$, we measure its impact by looking at the lth dimension's mean's relative error,

$$\phi_{\varepsilon,l} \triangleq \frac{|\mathbb{E}[W_{il} - Y_{ijl}]|}{\sqrt{\mathrm{Var}[Y_{ijl}]}} = \frac{|\mathbb{E}[\varepsilon\sqrt{l}]|}{\sqrt{\mathrm{Var}[\mathbb{E}[Y_{ijl}|\Psi_i]] + \mathbb{E}[\mathrm{Var}[Y_{ijl}|\Psi_i]]}} = \frac{\varepsilon\sqrt{l}}{\sqrt{0.25 + 4.25l}}.$$

To select the values of ε for the experiments, we consider an asymptotic relative error by taking $l \to \infty$: $\phi_\varepsilon \triangleq \lim_{l\to\infty}\phi_{\varepsilon,l} = \frac{\varepsilon}{\sqrt{4.25}}$. We control the value of ε such that ϕ_ε equals 0, 0.25 and 0.5. Similarly, as δ controls the marginal variance in $\mathbf{W}_i$, we study the relative error in variance:

$$\phi_{\delta,l} \triangleq \frac{|\mathrm{Var}[W_{il}] - \mathrm{Var}[Y_{ijl}]|}{\mathrm{Var}[Y_{ijl}]} = \frac{|4.25(1-\delta)l|}{0.25 + 4.25l},$$

and it follows that $\phi_\delta \triangleq \lim_{l\to\infty}\phi_{\delta,l} = |1-\delta|$. We choose $\delta = 0.75, 1$ or 1.25 to make $\phi_\delta = 0$ or 0.25. Finally, we make $\theta \in \{0, 0.5, 1\}$ to control the dependency in measures within $\mathbf{W}$.

In our experiment, we set $n = 100$ as the number of epochs and fix the number of replications per epoch at $r = 1{,}000$. All tests are performed with a FWER of $\alpha = 0.05$. For marginal hypothesis testing, we utilize the KS and AD tests with the error rate, $\alpha_M = \frac{d}{d+1}\alpha$. We perform copula tests based on $\hat{T}_1$, $\hat{T}_2$ and $\hat{T}_3$ with the error rate, $\alpha_J = \alpha/(d+1)$. To examine the effect of the dimensions to the performances of the tests, we run the experiments with dimension $d = 3$ and 10. Subsequently, the values of α_M and α_J depend on d. The critical value of each test statistic is computed according to the bootstrap resampling procedure described in Section 4 with bootstrap size $B = 10{,}000$.

As mentioned in Section 3, we modify the ecdf formula in (3) to prevent infinite values when performing the AD test. In detail, for the lth performance measure in the ith epoch, let $Y_{i(1)l} \leq Y_{i(2)l} \leq \cdots \leq Y_{i(r)l}$ be the order statistics of the simulation replications $\{Y_{ijl}\}_{j=1}^r$. We modify $\hat{G}_{il}$ so that $\hat{G}_{il}(y) = \frac{1}{r+1}$ if $y < Y_{i(1)l}$ and $\hat{G}_{il}(y) = \frac{r}{r+1}$ if $y > Y_{i(r)l}$. For $y \in [Y_{i(1)l}, Y_{i(r)l}]$, we first apply linear interpolation on the step function (3),

and then utilize the transformation $\Gamma(x) = \frac{r-1}{r+1}x + \frac{1}{r+1}$ to scale the probability into the range $[\frac{1}{r+1}, \frac{r}{r+1}]$. To summarize, we adopt the following estimator for G_{il}:

$$\hat{G}_{il}(y) = \frac{1}{r+1} + \frac{r-1}{r+1}\sum_{j=1}^{r-1} \mathbf{1}\left\{Y_{i(j)l} \le y < Y_{i(j+1)l}\right\}\left(\frac{j}{r}\frac{(Y_{i(j+1)l}-y)}{Y_{i(j+1)l}-Y_{i(j)l}} + \frac{j+1}{r}\frac{(y-Y_{i(j)l})}{Y_{i(j+1)l}-Y_{i(j)l}}\right).$$

With this change, we adopt $\mathbf{Z}_{ij}^{\mathbf{Y}} = \hat{\mathbf{V}}_{ij}$ instead of scaling $\hat{\mathbf{V}}_{ij}$ by $r/(r+1)$ as discussed in Section 4.

Table 1 presents the rejection rates of the marginal and copula tests under different ε, δ and θ values when the dimension d is set to 3. Additionally, we increment n and report the mean rejection rates, along with their standard errors in parentheses, based on 4,000 macro runs. Note that when $\varepsilon = 0$, $\delta = 1$ and $\theta = 0.5$, the distributions of $\mathbf{W}$ and $\mathbf{Y}$ match perfectly. We mark these cases with an asterisk in the second column of controls. Table 1(a) focuses on the marginal test using KS and AD statistics. The error rate allocated for the marginal test is $\alpha_M = \frac{d}{d+1}\alpha = 0.0375$. For the first set of comparisons, we fix $\delta = 1$ and $\theta = 0.5$ while changing the value of ε. In this case, the marginal variances and copulas of $\mathbf{W}$ and $\mathbf{Y}$ perfectly match. When $\varepsilon = 0$, the detection rates for both KS and AD remain stable for different choices of n, meeting the α_M Type-I error. As ε increases, the AD test generally exhibits a higher detection rate compared to the KS test, particularly demonstrating an advantage when n is small. However, in the second setting when $\varepsilon = 0$ and $\theta = 0.5$, both KS and AD test fail to achieve satisfactory detection rates when $\delta = 1 \pm 0.25$. In Table 1(b), detection rates of copula tests are compared when $\varepsilon = 0$ and $\delta = 1$. In this case, the marginal distributions of $\mathbf{W}$ and $\mathbf{Y}$ are identical, but their copula structures differ. The error rate allocated for the copula test is $\alpha_J = \frac{1}{d+1}\alpha = 0.0125$. When θ is set to 0.5, all three methods achieve acceptable Type-I error. When θ is away from 0.5, both $\hat{T}_1$ and $\hat{T}_2$ show high rejection rates when n is big. Specifically, both methods exhibit higher detection rates when $\theta = 1$ than when $\theta = 0$. When the dependence decreases, $\hat{T}_2$ tends to outperform $\hat{T}_1$ and vice versa when dependence increases. It also shows that when the null hypothesis is not true, both $\hat{T}_1$ and $\hat{T}_2$ methods are not reliable when n is small. The $\hat{T}_3$ method fails to achieve a satisfactory detection rate when the null hypothesis is incorrect. Moreover, it also requires some fine-tuning of the bandwidth h in the Gaussian kernel.

In Table 2 we show the detection rate of the multiple hypothesis testing with $n = 50$ or 100 across all combinations of ε, δ and θ. We use asterisks to indicate the cases where the marginal mean/variance or the copula structures of $\mathbf{W}$ and $\mathbf{Y}$ are identical. The row with all three identical parameters is highlighted in which case the null hypothesis holds and the rejection rate $\alpha = 5\%$. Suggested by the results in Table 1, the AD test outperforms the KS test and $\hat{T}_2$ is better than $\hat{T}_1$ in most of the cases. To demonstrate the potential performance differences with respect to the choice of test statistics, we present the best and worst combinations of the marginal and copula tests in Table 2, (a) KS and $\hat{T}_1$ and (b) AD and $\hat{T}_2$. Both combinations yield detection rate close to $\alpha = 0.05$ when all three parameters are aligned. When the marginal means differ significantly, namely $\varepsilon = 1.0308$, the detection rates of both approach one and are robust regardless of the number of epochs n. When the marginal means are mildly different ($\varepsilon = 0.5154$), the detection rates generally stay high, except when $\theta = 0.5$. For $\theta = 0.5$, the detection rates of both (a) and (b) decrease dramatically when $n = 50$ and the former shows rates below 0.8 even at $n = 100$. The sensitivity of the detection rate on n becomes far more pronounced when $\varepsilon = 0$. The detection rates of (b) drop to ~ 0.2 in some cases. Similar to the observations made for Table 1, both combinations struggle when ε and θ match while δ does not. For the case $\varepsilon = 0$, $\delta = 0.75$ and $\theta = 0.5$, multiple hypothesis testing slightly improves upon the detection rates of the marginal tests. Overall, the data in Table 2 indicates (b) exhibits better power than (a) when the system and simulated KPIs do not match in distribution.

Table 3 shows the detection rates of (a) and (b) when dimension d is set to 10. When all values of ε, δ, and θ are consistent, both combinations give reasonable Type-I error around $\alpha = 0.05$. Compared with the results in Table 2, when the marginal means do not match ($\varepsilon \neq 0$), the detection rates of both (a) and (b) increase when d is increased from 3 to 10. When the marginal means coincide ($\varepsilon = 0$), the detection rates decrease dramatically when $\theta = 0$, exhibit a slight increase when $\theta = 0.5$, and grow significantly when

Table 1: Detection rate of marginal and copula tests for different values of ε, δ and θ when $d = 3$.

controls		test	$n = 20$	$n = 40$	$n = 60$	$n = 80$	$n = 100$
$\delta = 1$, $\theta = 0.5$	$\varepsilon^* = 0$	KS	0.0367(0.0030)	0.0352(0.0029)	0.0340(0.0029)	0.0393(0.0031)	0.0375(0.0030)
		AD	0.0357(0.0029)	0.0333(0.0028)	0.0362(0.0030)	0.0372(0.0030)	0.0375(0.0030)
	$\varepsilon =$ 0.5154	KS	0.2028(0.0064)	0.3585(0.0076)	0.5327(0.0079)	0.6655(0.0075)	0.7850(0.0065)
		AD	0.2595(0.0069)	0.4572(0.0079)	0.6462(0.0076)	0.7883(0.0065)	0.8855(0.0050)
	$\varepsilon =$ 1.0308	KS	0.6575(0.0075)	0.9360(0.0039)	0.9940(0.0012)	0.9995(0.0004)	1.0000(0.0000)
		AD	0.7875(0.0065)	0.9795(0.0022)	0.9990(0.0005)	1.0000(0.0000)	1.0000(0.0000)
$\varepsilon = 0$, $\theta = 0.5$	$\delta = 0.75$	KS	0.0302(0.0027)	0.0338(0.0029)	0.0375(0.0030)	0.0455(0.0033)	0.0530(0.0035)
		AD	0.0170(0.0020)	0.0200(0.0022)	0.0257(0.0025)	0.0367(0.0030)	0.0510(0.0035)
	$\delta = 1.25$	KS	0.0532(0.0036)	0.0512(0.0035)	0.0555(0.0036)	0.0658(0.0039)	0.0750(0.0042)
		AD	0.0732(0.0041)	0.0805(0.0043)	0.0955(0.0046)	0.0995(0.0047)	0.1235(0.0052)

(a) Detection rate of marginal tests with allocated error rate $\alpha_M = 0.0375$.

controls		test	$n = 20$	$n = 40$	$n = 60$	$n = 80$	$n = 100$
$\varepsilon = 0$, $\delta = 1$	$\theta = 0$	$\hat{T}_1$	0.0008(0.0004)	0.0515(0.0035)	0.2552(0.0069)	0.5240(0.0079)	0.7422(0.0069)
		$\hat{T}_2$	0.1772(0.0060)	0.4760(0.0079)	0.7037(0.0072)	0.8682(0.0053)	0.9410(0.0037)
		$\hat{T}_3$	0.0220(0.0023)	0.0330(0.0028)	0.0483(0.0034)	0.0620(0.0038)	0.0848(0.0044)
	$\theta^* = 0.5$	$\hat{T}_1$	0.0107(0.0016)	0.0107(0.0016)	0.0147(0.0019)	0.0132(0.0018)	0.0152(0.0019)
		$\hat{T}_2$	0.0118(0.0017)	0.0135(0.0018)	0.0138(0.0018)	0.0105(0.0016)	0.0107(0.0016)
		$\hat{T}_3$	0.0127(0.0018)	0.0158(0.0020)	0.0130(0.0018)	0.0077(0.0014)	0.0095(0.0015)
	$\theta = 1$	$\hat{T}_1$	0.3887(0.0077)	0.7065(0.0072)	0.8835(0.0051)	0.9583(0.0032)	0.9880(0.0017)
		$\hat{T}_2$	0.2203(0.0066)	0.5687(0.0078)	0.7855(0.0065)	0.9095(0.0045)	0.9698(0.0027)
		$\hat{T}_3$	0.0177(0.0021)	0.0470(0.0033)	0.0725(0.0041)	0.1275(0.0053)	0.1640(0.0059)

(b) Detection rates of the copula tests with nominal error rate $\alpha_J = 0.0125$.

$\theta = 1$. This may be explained by that for $d = 10$, we allocate a bigger portion of FWER to the marginal tests. Overall, (b) maintains an advantage in power over (a) when the $\mathbf{W}$ and $\mathbf{Y}$ distributions do not match.

ACKNOWLEDGMENTS

This work is partially funded by the National Science Foundation Grant CAREER CMMI-2045400.

REFERENCES

Balci, O. and R. G. Sargent. 1981. "A Methodology for Cost-Risk Analysis in the Statistical Validation of Simulation Models". *Communications of the ACM* 24(4):190–197.

Berg, D. 2009. "Copula Goodness-of-fit Testing: An Overview and Power Comparison". *The European Journal of Finance* 15(7-8):675–701.

Biller, B., X. Jiang, J. Yi, P. Venditti and S. Biller. 2022. "Simulation: the Critical Technology in Digital Twin Development". In *2022 Winter Simulation Conference*, 1340–1355 https://doi.org/10.1109/WSC57314.2022.10015246.

Deheuvels, P. 1979. "La Fonction de Dépendance Empirique et Ses Propriétés. Un Test Non Paramétrique d'Indépendance". *Bulletins de l'Académie Royale de Belgique* 65(1):274–292.

dos Santos, C. H., A. T. Campos, J. A. B. Montevechi, R. de Carvalho Miranda and A. F. B. Costa. 2023. "Digital Twin Simulation Models: A Validation Method Based on Machine Learning and Control Charts". *International Journal of Production Research* 62(7):2398–2414.

Durante, F., J. Fernández-Sánchez, and C. Sempi. 2013. "A Topological Proof of Sklar's Theorem". *Applied Mathematics Letters* 26(9):945–948.

Genest, C., J.-F. Quessy, and B. Rémillard. 2006. "Goodness-of-fit Procedures for Copula Models Based on the Probability Integral Transformation". *Scandinavian Journal of Statistics* 33(2):337–366.

Genest, C. and B. Rémillard. 2008. "Validity of the Parametric Bootstrap for Goodness-of-fit Testing in Semiparametric Models". *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 44(6):1096–1127.

Table 2: Detection rates of family-wise tests for different values of ε, δ and θ when $d = 3$.

controls			KS+$\hat{T}_1$		AD+$\hat{T}_2$	
ε	δ	θ	$n = 50$	$n = 100$	$n = 50$	$n = 100$
0.0000*	0.75	0	0.1502(0.0057)	0.7600(0.0068)	0.5972(0.0078)	0.9445(0.0036)
0.0000*	0.75	0.5*	0.0488(0.0034)	0.0683(0.0040)	0.0352(0.0029)	0.0607(0.0038)
0.0000*	0.75	1	0.8255(0.0060)	0.9905(0.0015)	0.6925(0.0073)	0.9742(0.0025)
0.0000*	1*	0	0.1537(0.0057)	0.7540(0.0068)	0.6050(0.0077)	0.9433(0.0037)
0.0000*	1*	0.5*	0.0498(0.0034)	0.0527(0.0035)	0.0460(0.0033)	0.0478(0.0034)
0.0000*	1*	1	0.8215(0.0061)	0.9885(0.0017)	0.6937(0.0073)	0.9710(0.0027)
0.0000*	1.25	0	0.1688(0.0059)	0.7615(0.0067)	0.6210(0.0077)	0.9465(0.0036)
0.0000*	1.25	0.5*	0.0665(0.0039)	0.0897(0.0045)	0.0938(0.0046)	0.1320(0.0054)
0.0000*	1.25	1	0.8260(0.0060)	0.9890(0.0016)	0.7023(0.0072)	0.9700(0.0027)
0.5154	0.75	0	0.6382(0.0076)	0.9815(0.0021)	0.8530(0.0056)	0.9992(0.0004)
0.5154	0.75	0.5*	0.5467(0.0079)	0.8978(0.0048)	0.5857(0.0078)	0.9417(0.0037)
0.5154	0.75	1	0.9100(0.0045)	0.9988(0.0006)	0.8465(0.0057)	0.9985(0.0006)
0.5154	1*	0	0.5563(0.0079)	0.9567(0.0032)	0.8377(0.0058)	0.9972(0.0008)
0.5154	1*	0.5*	0.4517(0.0079)	0.7880(0.0065)	0.5547(0.0079)	0.8875(0.0050)
0.5154	1*	1	0.8875(0.0050)	0.9968(0.0009)	0.8237(0.0060)	0.9935(0.0013)
0.5154	1.25	0	0.5620(0.0078)	0.9605(0.0031)	0.8700(0.0053)	0.9978(0.0007)
0.5154	1.25	0.5*	0.4540(0.0079)	0.7905(0.0064)	0.6148(0.0077)	0.9060(0.0046)
0.5154	1.25	1	0.8798(0.0051)	0.9962(0.0010)	0.8340(0.0059)	0.9930(0.0013)
1.0308	0.75	0	0.9978(0.0007)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
1.0308	0.75	0.5*	0.9950(0.0011)	1.0000(0.0000)	0.9988(0.0006)	1.0000(0.0000)
1.0308	0.75	1	0.9975(0.0008)	1.0000(0.0000)	0.9988(0.0006)	1.0000(0.0000)
1.0308	1*	0	0.9935(0.0013)	1.0000(0.0000)	0.9995(0.0004)	1.0000(0.0000)
1.0308	1*	0.5*	0.9805(0.0022)	1.0000(0.0000)	0.9950(0.0011)	1.0000(0.0000)
1.0308	1*	1	0.9925(0.0014)	1.0000(0.0000)	0.9948(0.0011)	1.0000(0.0000)
1.0308	1.25	0	0.9872(0.0018)	1.0000(0.0000)	0.9990(0.0005)	1.0000(0.0000)
1.0308	1.25	0.5*	0.9673(0.0028)	0.9995(0.0004)	0.9922(0.0014)	1.0000(0.0000)
1.0308	1.25	1	0.9865(0.0018)	0.9998(0.0002)	0.9910(0.0015)	1.0000(0.0000)

Hua, E. Y., S. Lazarova-Molnar, and D. P. Francis. 2022. "Validation of Digital Twins: Challenges and Opportunities". In *2022 Winter Simulation Conference*, 2900–2911 https://doi.org/10.1109/WSC57314.2022.10015420.

Kleijnen, J. P. C., B. Bettonvil, and W. Van Groenendaal. 1998. "Validation of Trace-Driven Simulation Models: A Novel Regression Test". *Management Science* 44(6):812–819.

Lugaresi, G., G. Aglio, F. Folgheraiter, and A. Matta. 2019. "Real-time Validation of Digital Models for Manufacturing Systems: A Novel Signal-processing-based Approach". In *Proceedings of the 2019 IEEE 15th International Conference on Automation Science and Engineering*, edited by S. S. Reveliotis, D. Cappelleri, D. V. Dimarogonas, M. Dotoli, M. P. Fanti, P. LUTZ, C. Seatzu, and X. Xie, 450–455. New Jersey: Institute of Electrical and Electronics Engineers.

Lugaresi, G., S. Gangemi, G. Gazzoni, and A. Matta. 2023. "Online Validation of Digital Twins for Manufacturing Systems". *Computers in Industry* 150:103942.

Morgan, L. E. and R. R. Barton. 2022. "Fourier Trajectory Analysis for System Discrimination". *European Journal of Operational Research* 296(1):203–217.

Naylor, T. H. and J. M. Finger. 1967. "Verification of Computer Simulation Models". *Management Science* 14(2):B–92–B–101.

Oakley, D., B. S. Onggo, and D. Worthington. 2020. "Symbiotic Simulation for the Operational Management of Inpatient Beds: Model Development and Validation using Δ-method". *Health Care Management Science* 23:153–169.

Panchenko, V. 2005. "Goodness-of-fit test for copulas". *Physica A: Statistical Mechanics and its Applications* 355(1):176–182.

Rhodes-Leader, L. A. and B. L. Nelson. 2023. "Tracking and Detecting Systematic Errors in Digital Twins". In *2023 Winter Simulation Conference*, 492–503 https://doi.org/10.1109/WSC60868.2023.10408052.

Romano, J. P. and M. Wolf. 2005. "Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing". *Journal of the American Statistical Association* 100(469):94–108.

Sargent, R. G. 2015. "Model Verification and Validation". In *Modeling and Simulation in the Systems Engineering Life Cycle: Core Concepts and Accompanying Lectures*, edited by M. L. Loper, 57–65. London: Springer.

Schruben, L. W. 1980. "Establishing the Credibility of Simulations". *Simulation* 34(3):101–105.

Table 3: Detection rates of family-wise tests for different values of ε, δ and θ when $d = 10$.

controls			KS+$\hat{T}_1$		AD+$\hat{T}_2$	
ε	δ	θ	$n = 50$	$n = 100$	$n = 50$	$n = 100$
0.0000*	0.75	0	0.0498(0.0034)	0.0747(0.0042)	0.0217(0.0023)	0.9950(0.0011)
0.0000*	0.75	0.5*	0.0488(0.0034)	0.0767(0.0042)	0.0262(0.0025)	0.0490(0.0034)
0.0000*	0.75	1	1.0000(0.0000)	1.0000(0.0000)	0.9998(0.0002)	1.0000(0.0000)
0.0000*	1*	0	0.0480(0.0034)	0.0540(0.0036)	0.0475(0.0034)	0.9960(0.0010)
0.0000*	1*	0.5*	0.0467(0.0033)	0.0548(0.0036)	0.0437(0.0032)	0.0498(0.0034)
0.0000*	1*	1	1.0000(0.0000)	1.0000(0.0000)	0.9992(0.0004)	1.0000(0.0000)
0.0000*	1.25	0	0.0793(0.0043)	0.1115(0.0050)	0.1388(0.0055)	0.9960(0.0010)
0.0000*	1.25	0.5*	0.0770(0.0042)	0.1035(0.0048)	0.1163(0.0051)	0.1605(0.0058)
0.0000*	1.25	1	1.0000(0.0000)	1.0000(0.0000)	0.9992(0.0004)	1.0000(0.0000)
0.5154	0.75	0	0.8325(0.0059)	0.9982(0.0007)	0.8892(0.0050)	1.0000(0.0000)
0.5154	0.75	0.5*	0.6715(0.0074)	0.9583(0.0032)	0.6860(0.0073)	0.9705(0.0027)
0.5154	0.75	1	1.0000(0.0000)	1.0000(0.0000)	0.9992(0.0004)	1.0000(0.0000)
0.5154	1*	0	0.7398(0.0069)	0.9842(0.0020)	0.8845(0.0051)	1.0000(0.0000)
0.5154	1*	0.5*	0.5825(0.0078)	0.8872(0.0050)	0.6715(0.0074)	0.9365(0.0039)
0.5154	1*	1	1.0000(0.0000)	1.0000(0.0000)	0.9995(0.0004)	1.0000(0.0000)
0.5154	1.25	0	0.7490(0.0069)	0.9848(0.0019)	0.9283(0.0041)	1.0000(0.0000)
0.5154	1.25	0.5*	0.5910(0.0078)	0.8875(0.0050)	0.7370(0.0070)	0.9530(0.0033)
0.5154	1.25	1	1.0000(0.0000)	1.0000(0.0000)	0.9998(0.0002)	1.0000(0.0000)
1.0308	0.75	0	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
1.0308	0.75	0.5*	0.9990(0.0005)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
1.0308	0.75	1	1.0000(0.0000)	1.0000(0.0000)	0.9998(0.0002)	1.0000(0.0000)
1.0308	1*	0	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
1.0308	1*	0.5*	0.9958(0.0010)	1.0000(0.0000)	0.9990(0.0005)	1.0000(0.0000)
1.0308	1*	1	1.0000(0.0000)	1.0000(0.0000)	0.9998(0.0002)	1.0000(0.0000)
1.0308	1.25	0	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)	1.0000(0.0000)
1.0308	1.25	0.5*	0.9905(0.0015)	1.0000(0.0000)	0.9978(0.0007)	1.0000(0.0000)
1.0308	1.25	1	1.0000(0.0000)	1.0000(0.0000)	0.9998(0.0002)	1.0000(0.0000)

Shannon, R. E. 1976. *Systems Simulation: The Art and Science*. New Jersey: Prantice-Hall, Inc.

AUTHOR BIOGRAPHIES

LINYUN HE is a Ph.D. candidate in the School of Industrial and Systems Engineering at Georgia Institute of Technology. His research interests include simulation optimization, stochastic optimization, non-parametric methods and high-dimensional statistics. His email address is lhe85@gatech.edu and his website is https://he-linyun.github.io.

LUKE RHODES-LEADER is a Lecturer in Management Science at Lancaster University, UK. His research interests include applications of simulation optimization and methodological aspects of digital twins. His email address is l.rhodesleader@lancaster.ac.uk, and his website is https://www.lancaster.ac.uk/lums/people/luke-rhodes-leader.

EUNHYE SONG is a Coca-Cola Foundation Early Career Professor and Assistant Professor in the H. Milton Stewart School of Industrial and Systems Engineering at Georgia Institute of Technology. Her research interests include simulation design of experiments, uncertainty and risk quantification, and simulation optimization. Her email address is eunhye.song@isye.gatech.edu. Her website is http://eunhyesong.info.