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.

A HYBRID APPROACH COMBINING SIMULATION AND A QUEUEING MODEL FOR OPTIMIZING A BIOMANUFACTURING SYSTEM

Danielle F. Morey[1], Giulia Pedrielli[2], and Zelda B. Zabinsky[1]

[1]Dept. of Industrial and Systems Eng., University of Washington, Seattle, WA, USA
[2]School of Computing and Augmented Intelligence, Arizona State University, Tempe, AZ, USA

ABSTRACT

We explore a hybrid approach to designing a biomanufacturing system with low-volume, high variability, and individualized products. Simulating a large number of possible configurations to determine those that meet target production goals is computationally impractical. We create an explainable surrogate model, specifically a queueing network model, that is calibrated to the output of a few computationally expensive simulations. The queueing network model enables a quick exploration of large numbers of mixed integer-continuous configurations, which would be challenging for traditional surrogate-based approaches. The queueing network model is used to quickly identify promising regions where a few configurations can then be evaluated with the simulation. The difference in evaluations at these configurations is used to decide whether the queueing model requires partitioning and/or re-calibration. The use of this hybrid approach with an explainable surrogate enables analysis, such as identifying bottle-necks, and gives insight into robust designs of the biomanufacturing system.

1 INTRODUCTION

Biomanufacturing provides unique challenges for simulation and optimization. These applications tend to have low-volume production of highly variable personalized products. Production often requires timely coordination between operators and machines across multiple production phases. Additionally, the resulting yield is variable, which can cause the need for rework.

In addition to these structural challenges, these biomanufacturing applications can be difficult to model due to a lack of data. Simulation, built from subject matter expert knowledge, can be a powerful tool. These simulations, often discrete event simulations, are equipped to handle the many intricacies and randomness of the true system. Their complexity, however, makes them computationally expensive. In some cases, this can even lead to numeric issues, especially near edge of feasibility where the system may not perform as expected.

A common approach to address the issues that come with complicated simulations is to generate a surrogate model of the simulation. In fact, surrogate models represent a way to generate hybrid simulations, where high fidelity simulators are associated to specific input locations, while predictions produced by lower fidelity, computationally more efficient models are associated to a complementary set of inputs. One such common surrogate is the Gaussian process model (Valentine and Sambridge 2020; Schulz et al. 2018; Santner et al. 2003; Rasmussen and Williams 2006; Shen et al. 2005; Cressie 2015). Gaussian process models are quite generalizable and, as such, are effective at modeling a wide range of functions. With the advancement of machine learning, neural networks have also become a popular choice for surrogate modeling (Cen and Haas 2023; Li et al. 2022; Tripathy and Bilionis 2018; Tsay 2021; Zhang et al. 2022).

Both of these surrogate modeling methods, however, lack explainability. If we have some understanding of the structure of the system, we can utilize an explainable model to more effectively gain insight into designs with good performance. The modeling method developed in this research utilizes a Jackson network queueing model to represent the complex system while exploiting the properties of queueing systems. A

similar methodology has been used to analyze and predict behavior of a semiconductor wafer fabrication facility (Rose 2007).

An important component of surrogate modeling is calibration. The process of calibration is the estimation of model parameter values to fit a model to data. There are many methods of calibration spanning many different applications (Lee et al. 2019; Lee et al. 2023; Osborne 1991; Pernot and Cailliez 2017; Speich et al. 2021). The calibration problem seeks to identify a set of model parameters that most closely match model outputs to data. Several calibration methods seek to minimize a loss function, for example, minimize mean squared error, while other methods accept calibration parameters with associated outputs within targeted confidence intervals of output metrics.

In this research, we calibrate a sequential queueing model to the outputs of a discrete event simulation. We demonstrate the benefits of this hybrid approach of utilizing an explainable surrogate model in exploring the decision space and gaining insights into the most promising solutions. Namely, we explore the trade-off between the number of servers (i.e., operators and machines) and the expected wait time of the system. The Pareto optimal solution set was identified based on the results of the explainable surrogate model. This solution set was then validated by comparing a selection of the queueing model results to the simulation results and observing consistent trends.

The rest of this paper is organized as follows. Section 2 discusses a discrete event simulation of a biomanufacturing application. Section 3 then introduces a queueing model representation of the same system. Section 4 discusses the calibration of the queueing model outputs to the simulation outputs. Section 5 utilizes the calibrated queueing model to explore the most promising designs. The selected designs are then evaluated in the simulation to confirm their performance. Finally, Section 6 provides a summary and conclusions.

2 DISCRETE EVENT SIMULATION

The biomanufacturing application that is the topic of this paper is an individualized cancer treatment manufacturing system. This system first involves the harvesting of T-cells from patients. In parallel, virus is harvested via outside processes. In the event that the cells must be transported to another facility, additional steps are required for preservation. Then, viral transduction and cell expansion occurs before the cells induction back to the patient. The viral transduction and cell expansion portions are referred to in this paper as "processing" for brevity. This process results in a variable yield of usable cells, which is tracked by the simulation. In the event that the yield becomes too low, new T-cells must be harvested. This process is depicted in Figure 1. A discrete event simulation has been developed to represent the biomanufacturing system with tune-able input parameters (Liu 2023a; Liu 2023b; Sharma 2020). In this paper, the simulation is executed with all jobs being processed in the same facility and with sufficient required resources, including virus and reagents.

This biomanufacturing system generates a personalized product and, as such, is highly dependent on the patient for which the product is being made. As such, patient variability is a key source of uncertainty in the system. More specifically, some patients' cells will have high responsiveness to the processing and result in higher yields, while others will have low responsiveness and result in especially low yield.

3 QUEUEING MODEL

The procedural nature of the biomanufacturing systems lends itself quite naturally to a Jackson network queueing model representation, which is depicted in Figure 2. The stages of the simulation are consolidated into sequential M/M/s queueing systems, as indicated by the dashed red boxes in Figure 1 and the corresponding boxes in Figure 2. More specifically, the process steps relating to T-cell harvesting in the simulation are grouped into two sequential queues in the queueing model. The first of the queues is labeled as "Setup" and consists of the components requiring a human operator. The second queue is labeled "Work" and consists of the components only requiring machinery. Similarly, the viral transduction and

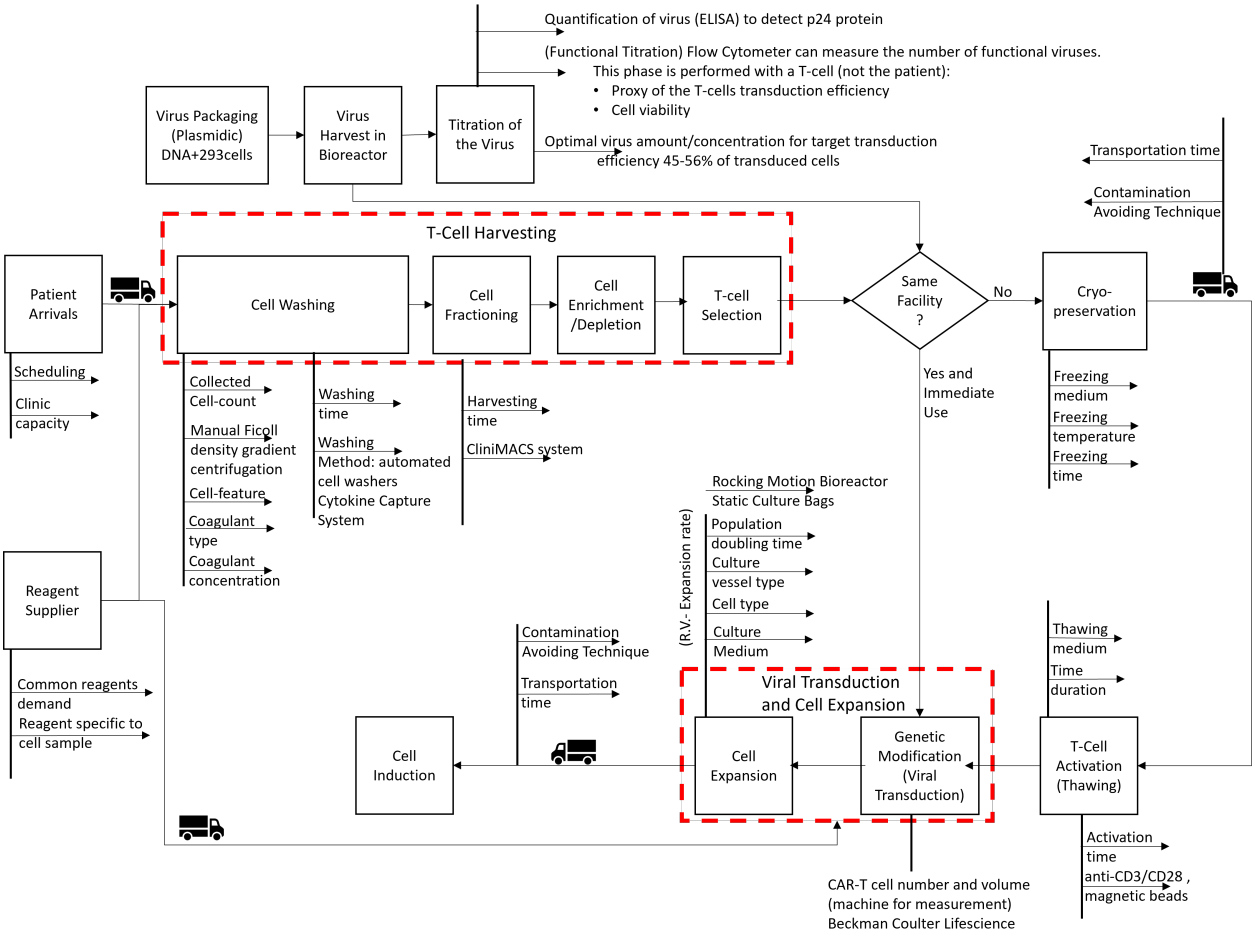

Figure 1: Flow chart of a biomanufacturing application.

cell expansion steps in the simulation, or processing, are consolidated and then split into a "Setup" and "Work" portion in the queueing model. In a sequential nature, the transition probability of going from one queue to the next in the queueing model, as indicated by the blue arrows in Figure 2, is 1 except in the case of the "Processing Work queue". In this stage, there is a non-zero probability that rework will be required due to low yield. To reflect this, there is a probability of $P(rework)$ to transition from "Processing Work" to "Harvesting Setup" and a resulting probability of $1-P(rework)$ to transition out of the system. Under the additional simplifying assumptions of exponentially distributed inter-arrival times and service rates, this Jackson network model has explainable parameters with real-world significance and can be solved analytically.

Some queueing model parameters were directly found from the simulation inputs while others require calibration. The simulation is set so that a new job enters the system with a randomly drawn inter-arrival time of either 1 or 2 days, or 1.5 days on average. In order to comply with the Jackson network queueing model assumptions, the inter-arrival time must be exponentially distributed. To match this in the queueing model, the mean arrival rate is set to $a_1 = 0.66$ arrivals per day, which results in the same average inter-arrival time of 1.5 days. Similarly, service times in the queueing model are assumed to be exponentially distributed. The simulation defines a range of potential service times for harvesting setup, harvesting work, and processing setup, and then draws one at random. The expected service time for each process in the simulation was then used as the expected service time for the exponential distribution in the queueing model, resulting in $t_1 = 2$ days, $t_2 = 7$ days, and $t_3 = 1$ day.

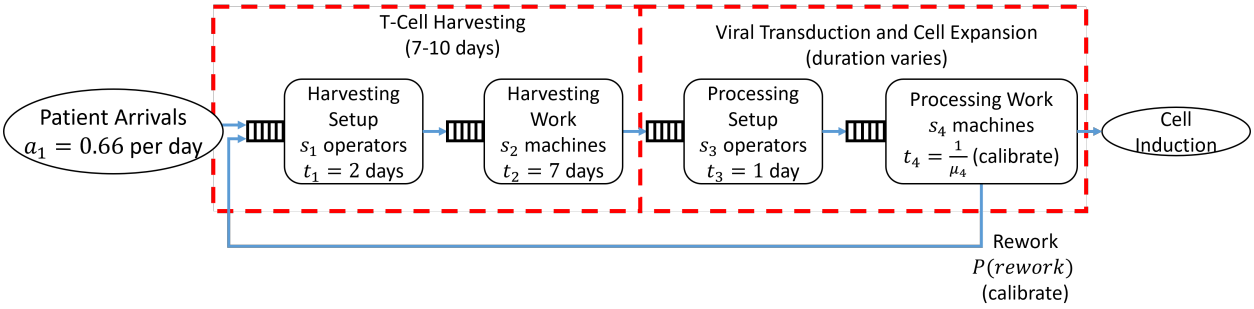

Figure 2: A Jackson network queueing model interpretation of the biomanufacturing system.

Some of the queueing model parameters have no direct representation in the simulation and thus must be calibrated. The processing work time, t_4, is dependent on the attributes of the individual patient, including age, sex, and blood volume, and thus varies. In the simulation, need for rework is determined by the yield, which is itself a function of many attributes, including patient responsiveness and processing duration. In the queueing network, rework is determined by a probability, $P(\text{rework})$, and thus that probability must be calibrated to reflect the simulation. These parameters, whether directly determined from the simulation or requiring calibration, are displayed in Figure 2.

Additional inputs to both the simulation and queueing model include the number of operators and machines for each of the processes, s_1, s_2, s_3, and s_4, jointly referred to as "servers" for brevity. These inputs are independent variables of both the simulation and the queueing model and are decided by the user.

Table 1: Parameter sets run by the discrete event simulation.

Parameter	**Levels**
s_1	$\{4,5\}$
s_2	$\{8,10\}$
s_3	$\{4,5\}$
s_4	$\{14,19\}$
Patient Mix	{Low Response, High Response}

4 MODEL CALIBRATION

Five parameters of the simulation were varied for analysis. Four of those are the independent variables of the number of servers, s_1, s_2, s_3, and s_4. The remaining parameter, called patient mix, is a representation of the uncertainty in how each individual patient's cells will respond to the process. Some patient's cells are quite responsive to the process, which results in higher yield, while other are less responsive, resulting in lower yield. The patient mix called "Low Response" is the case in which a higher proportion of patients' cells exhibit lower responsiveness, while the patient mix called "High Response" is the case in which a higher proportion of patients' cells exhibit higher responsiveness. This parameter is an unknown input of the system and it is thus important to explore both levels. Two levels are selected for each of these five parameters, as detailed in Table 1, resulting in $2^5 = 32$ different parameter sets.

The simulation was run for each of the 32 parameters sets with 40 replications, and the average time each job spent in the system was recorded. Figure 3 displays the average time in system across the 40 replications along with the 99% confidence interval.

The calibration parameters were determined by searching a Latin hypercube and down-selecting based on the resulting queueing outputs. A Latin hypercube sampling scheme was utilized to generate $100,000$ queueing calibration parameter combinations for $t_4 \in [1,100]$ and $P(\text{rework}) \in [0,1]$. For each of the 32

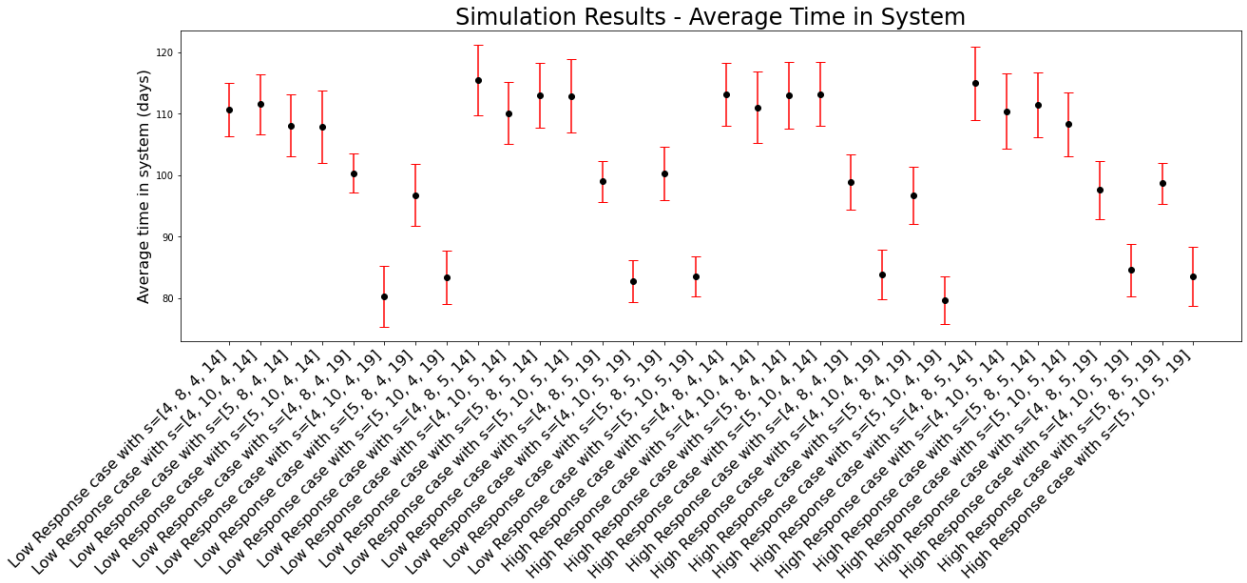

Figure 3: Simulation results for average time in system for each parameter set with 99% confidence interval.

simulation parameter sets, all 100,000 queueing calibration parameter combinations were input into the queueing model. Queueing calibration parameter combinations that resulted in an average time in system within the 99% confidence interval identified by the simulation for all 32 simulation parameter sets were accepted.

However, this initial calibration attempt was unsuccessful in the sense that there were no queueing calibration parameter combinations that generated queueing network outputs within the tolerance for all 32 simulation parameter sets. This indicated that the model needed to be improved before an accurate calibration could be found. By looking at Figure 3, it can be seen that cases in which $s_4 = 14$ resulted in approximately the same average time in the system. Cases in which $s_4 = 19$, on the other hand, varied based on the value of the other parameters. This suggested a natrual partition to the model on s_4. The model was thus improved by splitting the state space into two groups, such that one queueing model was be calibrated for parameter sets with $s_4 <= 16$ and a second queueing model for parameter sets with $s_4 > 16$. Using the same 100,000 Latin hypercube samples, one successful calibration was identified for parameter sets with $s_4 = 14$ and a different successful calibration was identified for parameter sets with $s_4 = 19$. The accepted values for the calibration parameters t_4 and $P(\text{rework})$ are detailed in Table 2.

Table 2: Summary of calibration results.

	s_1	s_2	s_3	s_4	Patient Mix	t_4	$P(\text{rework})$
Calibration A	{4,5}	{8,10}	{4,5}	{14}	{Low, Hi}	19.90066051	0.03617732
Calibration B	{4,5}	{8,10}	{4,5}	{19}	{Low, Hi}	17.04246784	0.38201367

5 IDENTIFY DESIGNS WITH GOOD PERFORMANCE

We aim to reduce average time in system with the fewest servers possible. In other words, we need to balance the number of servers with the average time in the system. This leads to a multi-objective black-box optimization problem where one objective is to minimize the total number of servers and another objective is to minimize the average time in the system. This problem, which requires many samples to identify a Pareto optimal solution set, quickly becomes intractable with the discrete event simulation alone. The calibrated queuing model can instead be leveraged to efficiently sample a large set of different server

number combinations and report the average time in the system. The set of non-dominated solutions, which is defined by the set of solutions where an improvement on one objective cannot be achieved without worsening on another objective, can then be analyzed by the simulation for validity.

A Latin hypercube sampling scheme was utilized to generate $10,000$ combinations of server sets for $s_1 \in [1,30]$, $s_2 \in [2,50]$, $s_3 \in [1,30]$, and $s_4 \in [2,30]$. Each server set was run for one of the two calibrated queueing models based on its value of s_4. More specifically, any server sets with $s_4 \leq 16$ used Calibration A while any server sets with $s_4 > 16$ used Calibration B, as given in Table 2. Any sets that result in queue utilization of over 1, indicating a queue that fails to reach steady-state, were discarded. The resulting queuing model outputs for average time in system are plotted in Figure 4 where the total number of servers, $s_1 + s_2 + s_3 + s_4$, is plotted on the horizontal axis. Many of the server sets tested could achieve the same or shorter average time in system for less total servers. The nondominated server sets, highlighted in green in Figure 4, reside on the efficient frontier between average time in system and total servers.

Figure 5 shows the first 15 nondominated server sets, which can be considered some of the most promising designs. These first few nondominated server sets can be considered the "knee" of the curve, as adding additional servers elicits very little additional improvement in average time in system. The results in Figure 5 indicate that 16 processing machines, s_4, and 2 or 3 processing operators, s_3, appear in most promising designs. When deciding on the optimal number of harvesting operators, s_1, and machines, s_2, it should be noted that there is a trade-off between the number of operators and machines and the resulting average time in system. More specifically, increased harvesting servers and machines will decrease the average time, but the marginal improvement decreases with each added server, as seen in Figure 5.

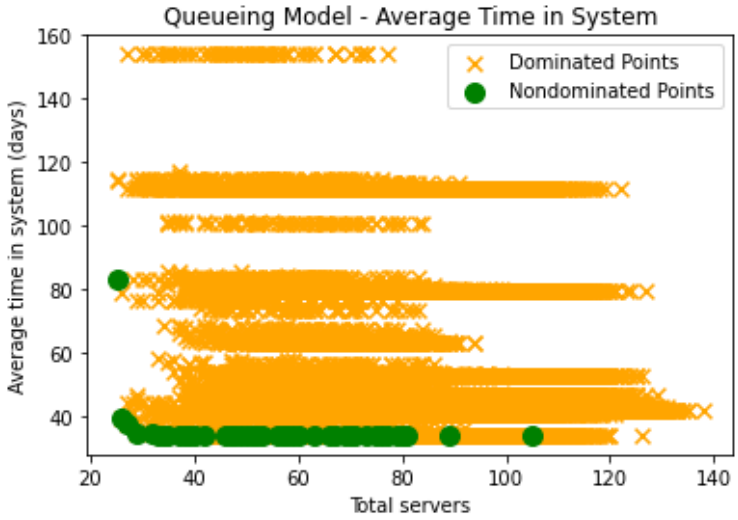

Figure 4: Queueing model results for average time in system across $10,000$ combinations of servers with nondominated points in green.

In order to validate the results elicited from the queueing model, four of the most promising server sets were also run in the simulation. The simulation results for both the low response case and the high response case, along with the queueing model results, are shown in Figure 6. While the simplifying assumptions in the queueing model lead to some inaccuracies in the output average time in system, the trends between the queueing model and the simulation remain consistent. This indicates that, while it is important to utilize the simulation to ensure accuracy, the queueing model is successful in identifying the trends in the system and can help to highlight promising designs for further analysis.

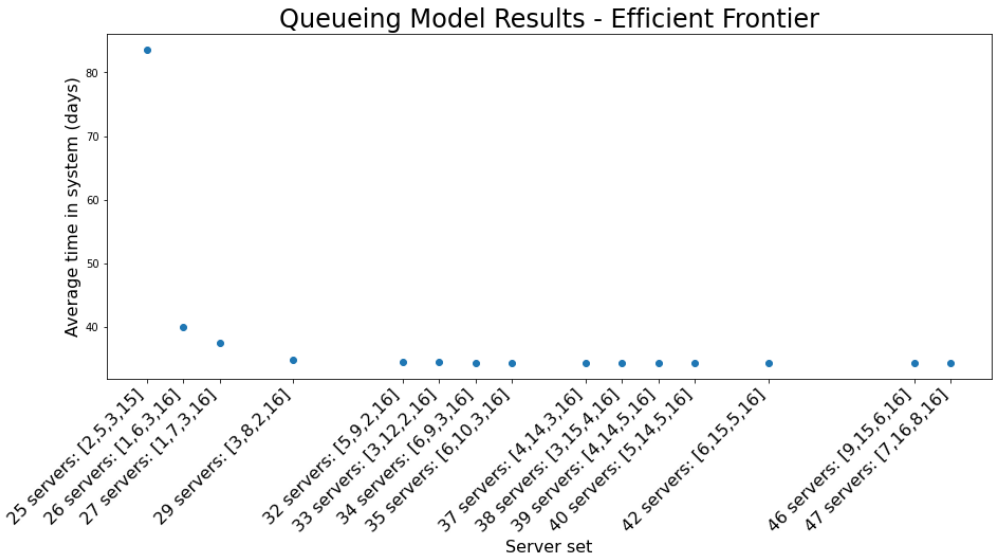

Figure 5: Details of the Pareto optimal solution sets, including time in system and their respective number of servers.

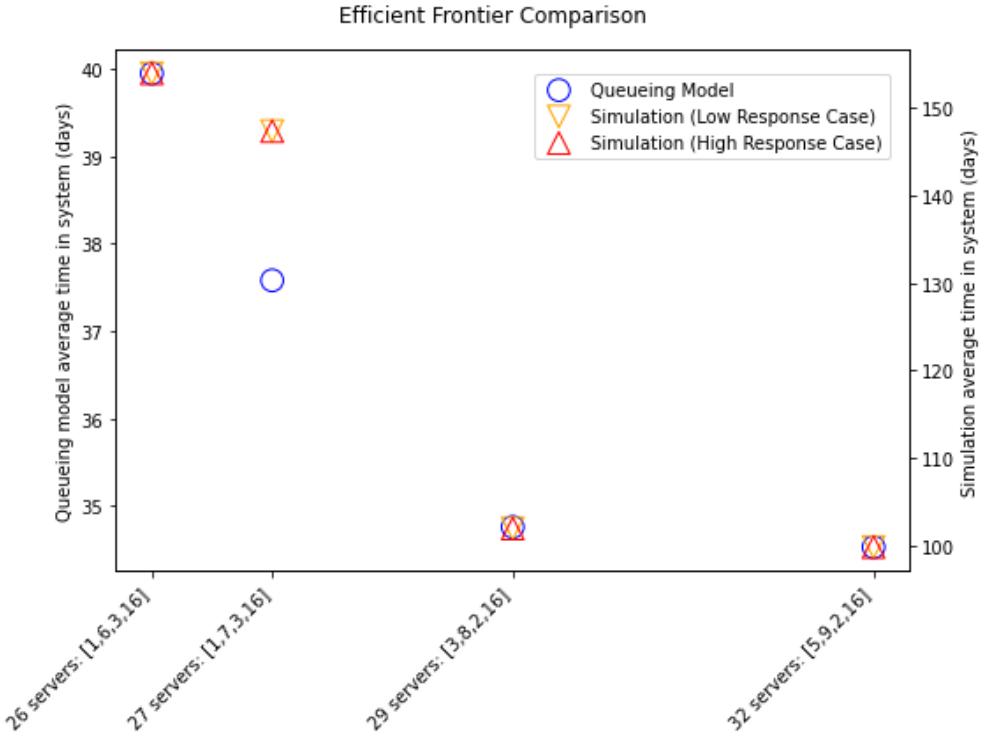

Figure 6: Comparison of simulation and queueing model results for 26-32 servers.

6 CONCLUSION

In this research, we used a hybrid approach consisting of calibrating a queueing model by utilizing the outputs of a discrete event simulation, to represent a biomanufacturing system. The explainability of the queueing model allowed for insights to be gained with regards to well-performing designs. This type

of study utilizing an explainable analytic model in conjunction with a simulation can be used in other applications besides biomanufacturing and with different explainable models other than a queueing model.

By analyzing the queueing model for a wide range of number of operators and machines we are able to identify the Pareto optimal solution set that minimizes average time in system and number of total operators and machines. The computationally cheap nature of the queuing model makes this analysis very easy to perform and the explainability further enhances our ability to draw conclusions from the results. Additionally, by performing this analysis on both the Low Response and High Response cases, we can further ensure that the system performs well under the uncertainty in the true system.

In future research, we aim to investigate other sources of uncertainty in the system, such as unexpected fluctuations in arrivals and fluctuations in manufacturing time, reflecting both stress and relaxed systems. Additionally, we intend to further explore the interplay between the simulation and the queueing model. This includes obtaining simulation results for more parameter sets and performing additional rounds of calibration to more closely examine how which parameters can be represented via the queueing model and which require separate calibrations.

REFERENCES

Cen, W. and P. J. Haas. 2023. "Efficient Hybrid Simulation Optimization via Graph Neural Network Metamodeling". In *2023 Winter Simulation Conference (WSC)*, 3541–3552 https://doi.org/doi:10.1109/WSC60868.2023.10408474.

Cressie, N. 2015. *Statistics for Spatial Dta, Revised Edition*. John Wiley & Sons.

Lee, G., W. Kim, H. Oh, B. D. Youn and N. H. Kim. 2019. "Review of statistical model calibration and validation - from the perspective of uncertainty structures". *Structural and Multidisciplinary Optimization* 60:1619–1644.

Lee, S., P. Maneekul, and Z. B. Zabinsky. 2023. "Representative Calibration Using Black-Box Optimization and Clustering". In *2023 Winter Simulation Conference (WSC)*, 3669–3680 https://doi.org/10.1109/wsc60868.2023.10408638.

Li, Z., F. Liu, W. Yang, S. Peng and J. Zhou. 2022. "A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects". *IEEE Transactions on Neural Networks and Learning Systems* 33(12):6999–7019.

Liu, M. 2023a. *Adaptive Gray Box Reinforcement Learning Methods to Support Therapeutic Research: From Product design to Manufacturing*. Ph.D. dissertation, Arizona State University.

Liu, M. 2023b. "Bio-Manufacturing-Simulator". GitHub. https://github.com/MenghanLiu212/Bio-Manufacturing-Simulator, accessed on 28th April, 2023.

Osborne, C. 1991. "Statistical Calibration: A Review". *International Statistical Review* 59(3):309–336.

Pernot, P. and F. Cailliez. 2017. "A critical review of statistical calibration/prediction models handling data inconsistency and model inadequacy". *American Institute of Chemical Engineers* 63(10):4642,4665.

Rasmussen, C. E. and C. K. Williams. 2006. *Gaussian processes for machine learning*, Volume 1. Cambridge, Massacnhusetts: the MIT Press.

Rose, O. 2007. "Improved simple simulation models for semiconductor wafer factories". In *2007 Winter Simulation Conference (WSC)*, 1708–171 https://doi.org/doi:10.1109/WSC.2007.4419793.

Santner, T. J., B. J. Williams, W. I. Notz, and B. J. Williams. 2003. *The design and analysis of computer experiments*, Volume 1. Springer.

Schulz, E., M. Speekenbrink, and A. Krause. 2018. "A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions". *Journal of Mathematical Psychology* 85:1–16.

Sharma, G. 2020. *Bioman: Discrete-event Simulator to Analyze Operations for Car-T Cell Therapy Manufacturing*. Ph.D. dissertation, Arizona State University. https://keep.lib.asu.edu/system/files/c7/233807/Sharma_asu_0010N_20167.pdf, accessed on 19th April, 2023.

Shen, Y., M. Seeger, and A. Ng. 2005. "Fast Gaussian process regression using KD-trees". *Advances in neural information processing systems* 18:1–8.

Speich, M., C. F. Dormann, and F. Hartig. 2021. "Sequential Monte-Carlo algorithms for Bayesian model calibration - A review and method comparison". *Ecological Modelling* 455.

Tripathy, R. K. and I. Bilionis. 2018. "Deep UQ: Learning deep neural network surrogate models for high dimensional uncertainty quantification". *Journal of Computational Physics* 375:565–588.

Tsay, C. 2021. "Sobolev trained neural network surrogate models for optimization". *Computers & Chemical Engineering* 153.

Valentine, A. P. and M. Sambridge. 2020, 3. "Gaussian process models-I. A framework for probabilistic continuous inverse theory". *Geophysical Journal International* 220:1632–1647.

Zhang, T., F. Li, X. Zhao, W. Qi and T. Liu. 2022. "A Convolutional Neural Network-Based Surrogate Model for Multi-objective Optimization Evolutionary Algorithm Based on Decomposition". *Swarm and Evolutionary Computation* 72.

AUTHOR BIOGRAPHIES

DANIELLE F. MOREY is a Ph.D. student in Industrial and Systems Engineering at University of Washington under the supervision of Zelda B. Zabinsky. She obtained her B.S. in the same field from The Ohio State University. Her research interests include multi-fidelity and multi-objective optimization. Her email address is dmorey43@uw.edu.

GIULIA PEDRIELLI is Associate Professor in the School of Computing and Augmented Intelligence at Arizona State University. She is interested in the area of stochastics and simulation based optimization, and deals with applications in biomanufacturing, power systems, supply chains, safety critical systems. Her email address is giulia.pedrielli@asu.edu.

ZELDA B. ZABINSKY is a Professor in the Department of Industrial and Systems Engineering at the University of Washington. She is an INFORMS Fellow and an IISE Fellow. Her Ph.D. is from the University of Michigan, in Industrial and Operations Engineering. Her research interests are in global optimization under uncertainty for complex systems, and she has worked in many application areas. Her email address is zelda@uw.edu. Her website is https://ise.washington.edu/facultyfinder/zelda-zabinsky.