Bridging Simulation and Practice: Surrogate-Based Production Scheduling for Stochastic Modular Construction

Nima Moghimi[1*], Sahar H. Shamaee[2], Qipei Mei[3], Vicente A. Gonzalez[4], Farook Hamzeh[5]

[1]Corresponding Author, Ph.D. Student, Department of Civil and Environmental Engineering, University of Alberta, Canada
[2]Graduate Research Assistant, Department of Civil and Environmental Engineering, University of Alberta, Canada
[3]Assistant Professor, Department of Civil and Environmental Engineering, University of Alberta, Canada
[4] Hal Kvisle Professor and Tier 1 Canada Research Chair in Digital Lean Construction, Infrastructure Human Tech (IHT) Lab, Strategic Projects Insight Centre in Engineering (SPICE), Department of Civil and Environmental Engineering, University of Alberta, Canada
[5]Professor, Department of Civil and Environmental Engineering, University of Alberta, Canada
nmoghimi@ualberta.ca, hamedsha@ualberta.ca, qipei@ualberta.ca, vagonzal@ualberta.ca, hamzeh@ualberta.ca

Abstract

In space-constrained modular construction, stochastic work content renders deterministic scheduling ineffective, often leading to production deadlocks. While Simulation-Based Optimization (SBO) addresses these dynamics, it typically lacks the operational utility for routine daily planning due to its high computational intensity. To bridge this gap, this study adopts a Design-Science Research (DSR) approach to design and evaluate a Surrogate-Based Scheduling Framework. The methodology follows a structured cycle where a Hierarchical Stochastic Demand Model generates a wide spectrum of synthetic production scenarios, which are evaluated via high-fidelity simulation to determine optimal Inter-Project Time Buffers. These complex optimization results are subsequently instantiated into a lightweight planning rule via surrogate-based heuristic derivation. To validate the utility of this artifact, a case study was conducted on a wood-frame wall panel assembly line. The resulting "Peak Workload" heuristic (instantiated via Linear Regression) achieved an R2 of 0.657, demonstrating that a simple linear approximation sufficiently captures dominant variability trends despite the system's inherent non-linearity. Operationally, the heuristic reduced average Cycle Time by 37.7% compared to the baseline while maintaining equivalent throughput. These results confirm the feasibility of transforming complex buffer optimizations into transparent, data-driven policies that stabilize production flow without requiring real-time computation.

Keywords – Modular Construction; Simulation-Based Optimization; Production Scheduling; Surrogate Modelling; Data-Driven Heuristics; Work-in-Process (WIP)

1 Introduction

Modular construction offers a promising, factory-based alternative to the inefficiencies of traditional building practices; however, its full potential is often constrained by unique scheduling complexities [1]. The Engineer-to-Order (ETO) nature of modular projects means that each component has a distinct work content, introducing significant process time variability [2]. In space-constrained production facilities, this unpredictability creates a precarious planning environment. Schedules that fail to account for this stochasticity often cause production lines to become either deadlocked with excess inventory or idle from a lack of available work, undermining the very efficiency the factory setting is meant to provide [3].

To navigate this uncertainty, production planners rely on look-ahead scheduling to sequence upcoming projects. However, the foundational tools for this task, such as the Critical Path Method (CPM), are inherently deterministic and ill-suited for dynamic environments [4,5]. Planners often compensate for this rigidity by inserting intuitive temporal buffers between projects [6]. This manual approach to creating "safety time" is unreliable; an insufficient gap between projects risks system-wide congestion, while an overly generous one leads to poor resource utilization and reduced capital efficiency [7,8]. The central challenge, therefore, is to develop a data-driven, simulation-informed optimization procedure that determines project-specific inter-project buffer times and is explicitly evaluated under stochastic variability [9,10].

Academic research has explored advanced computational methods to address this. Discrete Event Simulation (DES) has been used to model factory dynamics, while metaheuristic algorithms have been applied to optimize production sequences [11]. A disconnect persists, however, between these sophisticated optimization techniques and the practical cadence of production planning. While Simulation-Based Optimization (SBO) offers high rigorous validity, it often lacks the operational utility necessary for the routine generation of weekly schedules [12–14]. The operational reality of construction management rarely accommodates the computational expense or specialized expertise required to maintain complex optimization engines. Consequently, there is a critical need for a new class of decision-support artifacts that bridge this gap—providing the rigor of simulation without the computational burden [15].

To address this challenge, this study adopts a Design-Science Research (DSR) approach to design and evaluate a novel scheduling artifact: the Surrogate-Based Scheduling Framework. The objective is to equip planners with a scientific basis for determining project start times. The core logic of this artifact is that robust schedules must be derived from a realistic understanding of operational uncertainty. To achieve this, the framework utilizes a multi-level probabilistic demand model to generate a wide spectrum of potential production scenarios. Through extensive simulation, an optimization process identifies the ideal inter-project lag time needed to absorb variability. This complex relationship is then instantiated into a computationally simple planning rule via a Surrogate-Based Heuristic Derivation strategy. The final output is a practical decision-making tool that translates project-specific data into scientifically sized buffers, enabling the creation of resilient look-ahead plans without the need for real-time computational analysis.

2 Research Methodology

2.1 Research Approach

This study adopts the Design-Science Research (DSR) methodology [16]. DSR is a problem-solving paradigm focused on the creation and rigorous evaluation of innovative "artifacts" to address practical organizational challenges.

Consistent with the DSR process, this research follows a three-stage cycle:

1. Problem Identification (Relevance): Defining the need for a dynamic buffer allocation method that overcomes the limitations of static scheduling practices.
2. Artifact Design: Developing the "Offline Learning Framework." This framework operates through a structured learning cycle (Figure 1): (1) synthetic demand scenarios are generated to capture operational variability; (2) these scenarios are evaluated within a high-fidelity simulation model; (3) an optimization algorithm solves for the optimal Inter-Project Time Buffer (τ^*); and (4) the results are instantiated into a lightweight planning rule via surrogate-based heuristic derivation.
3. Evaluation (Rigor): Validating the utility of the artifact through a case study application (Section 3).

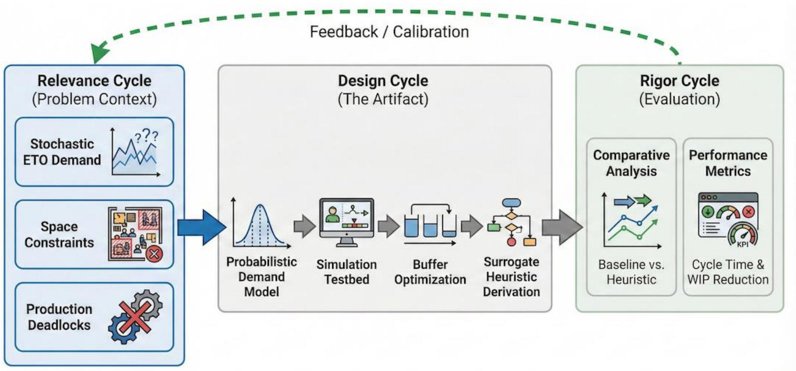

Figure 1: Structure of the proposed framework.

Unlike deterministic planning that assumes fixed task durations, this DSR artifact recognizes that robust schedules must be informed by the probabilistic reality of ETO production. The methodology shifts the decision point to the planning phase, allowing buffers to be sized scientifically before production begins.

2.2 Hierarchical Stochastic Demand Modelling

To ensure the scheduling artifact addresses the 'Relevance' requirement of DSR, the design process begins by modelling the environment's true variability. Therefore, a Hierarchical Stochastic Model synthesizes look-ahead windows that reflect the variability inherent in actual ETO demand. This scenario-based modelling logic is consistent with prior work that uses Monte Carlo / scenario sampling to generate diverse realizations of uncertain future conditions before evaluating candidate strategies [17,18]. The model integrates four interdependent sub-components:

Project Occurrence Process: The arrival of incoming orders varies with market conditions. The number of projects N_{proj} within a planning horizon is modeled as a Poisson process:

$$P(N_{proj} = k) = \frac{\lambda^k e^{-\lambda}}{k!} \quad (1)$$

where λ is the mean arrival rate, calibrated to reflect Low, Medium, or High workload intensities.

Project Scope: Each project has a defined work volume. The total quantity of units Q_i for Project *i* follows a discrete uniform distribution bounded by facility constraints:

$$Q_i \sim \mathcal{U}\{Q_{min}, Q_{max}\} \quad (2)$$

Product Mix (Complexity): Engineer-to-Order environments involve heterogeneous products with different complexity profiles. Let ρ denote the complexity ratio—the fraction of "complex" (high-variability) units within a project. The number of complex units is:

$$n_{c,i} = \text{Round}(\rho \cdot Q_i) \quad (3)$$

This parameter directly influences downstream processing time variability and resource contention, making it a critical input for buffer determination.

Delivery Constraints: External logistics (e.g., trailer availability, site readiness) constrain when projects can start. The allowable lead time δ_i is modeled stochastically to reflect real-world scheduling pressures:

$$D_i = T_{entry,i} + \text{Duration}_{est,i} + \delta_i \quad (4)$$

This temporal coupling ensures that the optimization considers both internal production variability and external logistical constraints.

2.3 Production System Simulation Environment

The simulation environment serves as the design testbed, allowing the artifact to be iteratively tested against non-linear dynamics before deployment. The stochastic production scenarios generated in Section 2.2 are evaluated within a computational simulation environment acting as a virtual testbed. While the specific modelling paradigm (e.g., Discrete Event Simulation, System Dynamics, or Agent-Based Modelling) may vary based on the facility's characteristics, the environment must faithfully replicate the non-linear dynamics of constrained production. The simulation framework generally operates on two critical layers to ensure validity:

Constraint-Based Physical Layer: The flow of work units is governed by rigorous physical and logical constraints rather than deterministic lead times. This includes finite buffer capacities (B_{Cap}) between workstations and specific resource availability windows (R_{avail}). By enforcing these limits, the simulation allows complex system behaviours—such as blocking (upstream work halting due to downstream congestion) and starvation (downstream idleness due to upstream delays)—to emerge endogenously from the system state rather than being pre-programmed.

Stochastic Process Execution: Task durations are modelled as stochastic variables drawn from probability distributions calibrated to historical performance data. To capture the heterogeneity of Engineer-to-Order (ETO) production, processing time P_t for a unit u is defined as a function of its specific complexity attributes:

$$P_t(u) = \begin{cases} f_{std}(\cdot) & \text{if } u \text{ is Standard Class} \\ f_{cplx}(\cdot) & \text{if } u \text{ is Complex Class} \end{cases} \quad (5)$$

where $f(.)$ represents a probability density function (e.g., Triangular, Log-Normal) appropriate for the specific task. This structure ensures that the optimization process specifically accounts for the volatility introduced by product mix variability, a primary driver of schedule unreliability in modular construction.

2.4 Optimization Formulation

This component represents the search process of the design cycle, identifying optimal configurations τ^* that maximize the utility function $J(\tau)$. The central problem is to determine the optimal Inter-Project Time Buffer vector $\tau^* = \{\tau_{1,2}, \tau_{2,3}, \dots, \tau_{N-1,N}\}$ that balances competing objectives. Here, $\tau_{i,i+1}$ denotes the scheduled lag between the planned completion of Project J_i and the start of Project J_{i+1}.

To avoid a highly fragmented ("scattered") schedule, the optimization procedure minimizes a weighted multi-objective cost function $J(\tau)$, which trades off process efficiency, production throughput, and overall schedule compactness:

$$J(\tau) = w_1 \cdot \widehat{CT} - w_2 \cdot \widehat{TH} + w_3 \cdot \widehat{MS} \quad (6)$$

where $\widehat{CT}$, $\widehat{TH}$, and $\widehat{MS}$ are the min-max normalized values of Cycle Time (CT), System Throughput (TH), and Makespan (MS), respectively. For a generic metric M, the normalized quantity is defined as:

$$\widehat{M} = \frac{M(\tau) - M_{min}}{M_{max} - M_{min}} \quad (7)$$

Within this framework, minimizing CT promotes efficient flow at the individual project level, while maximizing TH (achieved via the negative term) incentivizes higher production volume. The Makespan term (MS) serves as a penalty against schedule scattering, compelling the optimizer to select the tightest feasible buffers that avoid overlap without creating unnecessary idle time. The weights w_1, w_2, w_3 are user-specified and satisfy $w_1 + w_2 + w_3 = 1.0$.

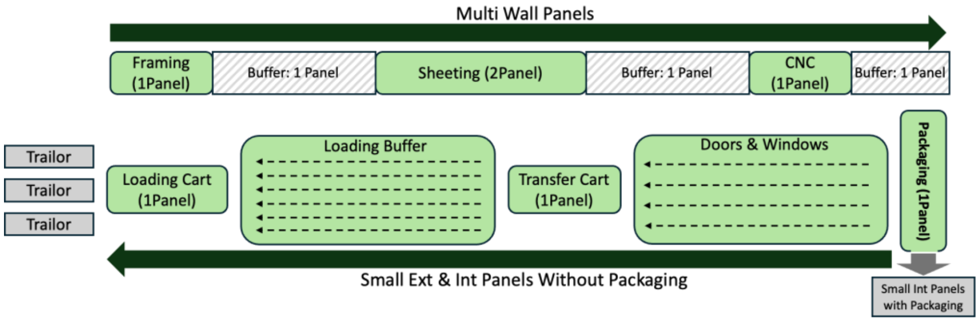

Figure 2: Schematic layout of the wall-assembly line.

The optimization is carried out under the domain constraint $0 \leq \tau_{i,i+1} \leq \tau_{max}$, where τ_{max} is the maximum commercially acceptable idle period (for example, one shift), ensuring that the solution is operationally realistic. The start-time relationship is given by:

$$Start_{i+1} = Finish_i^{planned} + \tau_{i,i+1} \quad 8$$

By solving this optimization across all synthetic scenarios, the algorithm produces a rich dataset of optimal buffer configurations, each linked to the project characteristics that generated it.

2.5 Surrogate Heuristic Derivation

The final step in the design cycle is to instantiate the complex optimization results into a usable tool. As running the full simulation-optimization loop for every scheduling update is computationally prohibitive for routine operations. To bridge the gap between high-fidelity analysis and practical daily planning, the framework synthesizes the complex relationships found by the optimization engine into a lightweight Planning Heuristic via surrogate modelling. This step effectively "crystallizes" the intelligence of the simulation into a static format.

We posit a functional relationship between the project state vector X and the optimal inter-project buffer τ^*. Crucially, X comprises not only independent attributes of the incoming project (e.g., work content Q_i) but also transition metrics that capture the shift in operational intensity between sequential projects (e.g., Δ Complexity or Δ Volume between J_i and J_{i+1}). This ensures the heuristic accounts for sequence-dependent setup risks.

During the design phase, the optimization algorithm generates a training dataset $D = \{(X_n, \tau_n^*)\}_{n=1}^{M}$ spanning M diverse scenarios. A regression analysis is performed on D to approximate the function f. While the specific architecture of the surrogate model can range from non-linear approximators to decision trees depending on data complexity, a linear approximation is often preferred for its interpretability in industrial contexts. The general form of the heuristic is defined as:

$$\tau^* \approx \beta_0 + \sum_{j=1}^{k} \beta_j \cdot x_j \quad 9$$

where, β_0, β_j are the learned coefficients representing the weight of each factor. x_j represents the Significant Predictors extracted from the project state vector X. These features are selected based on their correlation with production variability (e.g., total work content, resource intensity variability, or sequence-dependent complexity shifts).

The resulting output is a transparent, closed-form equation that eliminates the need for repeated simulation. This allows planners to implement the scientifically derived policy using standard spreadsheet tools or elementary calculation tables. By simply inputting the scheduled project attributesx_j, the heuristic instantaneously provides the optimal "safety time" τ^* required to maintain flow, transforming a computationally expensive optimization problem into a practical arithmetic task.

3 Artifact Evaluation: Case Study Implementation

To validate the proposed scheduling framework, a comprehensive case study was conducted within a real-world manufacturing environment. This section details the experimental setup, bridging the theoretical methodology with the specific physical constraints of an operating facility.

3.1 Evaluation Domain: Edmonton Assembly Line

The validation domain is a wood-frame wall panel assembly line in Edmonton, Alberta, organized as a flexible flow shop. As schematically illustrated in **Error! Reference source not found.**, the process flows sequentially from Framing through Sheeting and Nailing to Packaging. At the Packaging station, the flow splits: completed interior panels route directly to storage, while exterior panels proceed to four parallel

Door & Window (DW) installation lanes before reaching the Loading buffer. The layout is characterized by severe spatial constraints, featuring minimal buffer capacities (often limiting WIP to a single panel) between stations—a typical challenge in "volumetric" construction where large product dimensions restrict storage.

This specific facility was selected because it presents a convergence of internal variability and external logistical constraints that render static scheduling ineffective. The production environment is highly stochastic, with heterogeneous project specifications driving variable processing times. Crucially, the system is constrained by outbound logistics: panels must be loaded onto project-dedicated trailers, which have limited daily availability (avg. 3/day). In this tight spatial environment, traditional scheduling strategies that fail to account for variability frequently lead to immediate blocking when trailers are unavailable, causing cascading flow stagnation. This susceptibility to gridlock makes the line an ideal testbed for validating the proposed buffer optimization approach.

3.2 Configuration of the Simulation Testbed

To represent the stochastic production dynamics of the Edmonton facility, a hybrid simulation model was developed in Python using SimPy. The model combines a Discrete Event Simulation (DES) layer for material flow and capacity constraints with an Agent-Based Simulation (ABS) layer for workforce behavior. Physical constraints such as finite buffer capacities and the single-table Packaging limit are explicitly modeled, allowing blocking and starvation to emerge naturally from congestion. The ABS layer captures crew collaboration, non-linear productivity effects, and task-priority rules, with complex exterior panels prioritized over standard interior units.

The model was parameterized using empirical field data from the partner facility, including 100–150 processing-time records per station, supported by expert input to confirm logic and variability ranges. Processing times were differentiated by panel characteristics, such as interior versus exterior units. Triangular distributions were used as a practical representation of uncertainty in this data-constrained setting because they preserve minimum, most-likely, and maximum values and are easy to calibrate. The model was validated against a historical dataset of 84 Multi-Wall projects, achieving R^2 values of 0.95 for Framing and 0.89 for Doors & Windows, with MAPE between 16% and 18%.

Key assumptions include fixed observed routing, finite buffer capacities, rule-based crew collaboration and prioritization, station-specific triangular processing-time distributions, and schedule control through inter-project buffers rather than dynamic resequencing.

3.3 Evaluation Protocol

The framework was implemented through the 'Surrogate-Based Derivation' phase specified in the methodology. Leveraging the Hierarchical Stochastic Model, thirty distinct production scenarios were generated to encompass a broad range of workload intensities, spanning from low-demand periods (Poisson $\lambda = 8$) to high-intensity surges ($\lambda = 25$). The choice of 30 scenarios reflects a balance between coverage and computational tractability for this conceptual design stage. Rather than serving as an exhaustive sample, these scenarios were constructed to span low-, medium-, and high-intensity workload regimes and to expose the framework to diverse operating conditions. Moreover, because each scenario yields multiple adjacent project-pair decisions, the resulting surrogate was calibrated on approximately 800 optimized decision points rather than on only 30 scenario-level observations. While this was sufficient to derive and interpret a stable first-order heuristic, future work should examine convergence of the surrogate coefficients and scheduling performance as the number of generated scenarios increases.

For each scenario, the Simulation-Based Optimization engine determined the optimal Inter-Project Time Buffer (τ^*). The decision variable was the elapsed time (in minutes) between the planned finish of Project J_i and the release of Project J_{i+1}. The Genetic Algorithm minimized a weighted cost function that integrated Cycle Time, System Throughput, and Makespan. For this study, the weighting configuration was set to w_1=0.35 (Cycle Time), w_2=0.35 (Throughput), and w_3=0.30 (Makespan). Equal priority was assigned to the efficiency metrics (CT and TH), while the Makespan term served primarily as a 'control variable' to prevent the optimizer from suggesting an overly scattered schedule. It is important to note that this cost function is modular; in practice, facility managers can adjust these weights or introduce new metrics to reflect shifting operational priorities (e.g., prioritizing maximum volume during peak seasons).

3.4 Artifact Instantiation: Heuristic Calibration

To translate the simulation-based optimization results into a practical planning tool, a surrogate model was calibrated using the ~800 optimal decision points generated by the simulation. Linear Regression was selected for its interpretability and stability, with the aim of relating high-level project attributes to the optimal inter-project buffer (τ^*).

To identify the most informative predictors, a Theory-Guided Sparse Model Discovery approach was applied. The candidate library included raw production metrics (e.g., Multi-Wall count, Small Panel count) and derived workload-intensity terms. Using LASSO with bootstrap-based stability selection, the process filtered out weak predictors and showed that raw unit counts alone were insufficient. Instead, the model consistently retained variables linked to bottleneck workload intensity, particularly the peak boundary intensity between adjacent projects and the directional change in workload.

Based on these results, the final regression was defined as:

$$\tau^* = \alpha + \beta\, M_i + \gamma\, \Delta M_i^+ + \epsilon \qquad 10$$

Where:

$M_i = \max\,(\text{mode}_{i-1}, \text{mode}_i)$ signifies the Peak Workload Intensity across the two adjacent projects.

$\Delta M_i^+ = \max\,(\text{mode}_i - \text{mode}_{i-1}, 0)$ signifies the Anticipatory Workload Increase, triggered only when the incoming project is more demanding than the outgoing one.

The model was calibrated against the simulation-generated dataset, producing the following coefficients: $\alpha = -46.57$, $\beta = 0.186$, and $\gamma \approx -0.084$.

The coefficients illuminate distinct operational mechanisms. The positive weight on M_i ($\beta = 0.186$) reinforces that the primary constraint is the magnitude of the bottleneck; heavier project pairs require larger buffers. However, the negative coefficient for the anticipatory term ($\gamma \approx -0.084$) introduces a crucial insight: the "Clean-Slate Effect."

Mathematically, this term diminishes the required buffer when progressing from a lighter to a heavier project. Operationally, this suggests that a lighter outgoing project clears the system efficiently, establishing a stable, low-congestion environment for the subsequent demanding project. Conversely, a heavy outgoing project leaves residual variability in its wake, necessitating a larger safety margin for the next task regardless of its intensity.

The calibrated heuristic achieved an In-Sample Mean Absolute Error (MAE) of 32.46 minutes and a Root Mean Square Error (RMSE) of 37.40 minutes, with an R^2 of 0.657. Although an MAE of approximately 32 minutes indicates a noticeable gap from the theoretical optimum, its practical acceptability in this study is supported by the downstream validation results rather than by the error magnitude alone. When applied to the validation-week schedule, the calibrated heuristic still produced substantial system-level improvements, including a 37.7% reduction in average cycle time while maintaining statistically equivalent throughput. This suggests that the surrogate captured the dominant scheduling signal required for look-ahead decision support, even if individual buffer predictions were not exact. The R^2 value confirms that nearly two-thirds of the variability in optimal buffer sizing is captured by this transparent "Peak Workload" heuristic, establishing its validity as a dependable planning rule. However, the present study did not explicitly quantify how surrogate prediction error propagates into secondary operational risk metrics such as blocking frequency, due-date deviation, or extreme-case instability under peak conditions; this should be examined in future work.

It is acknowledged that the underlying production system exhibits inherent non-linear dynamics due to finite buffering and complex resource interactions. However, at this preliminary stage, high-fidelity feature data (e.g., granular task-level specifications) is often unavailable during the look-ahead planning phase. Consequently, a Linear Regression model was selected to demonstrate the potential of the methodology using only high-level metrics available to planners (e.g., panel counts, expected execution times). While an R^2 of 0.657 indicates that roughly one-third of the variance remains unexplained—suggesting the need for future non-linear modelling (e.g., Random Forests) as data granularity improves—the current model successfully captures the dominant 'Peak Workload' trend required to stabilize the schedule.

4 Results and Discussion

To assess the operational utility of the designed artifact, the methodology was evaluated using a calibrated simulation testbed initialized with a real 14-project production schedule from the partner facility. While a comprehensive live field implementation remains a subject for future research, this validation horizon was selected to align with the standard 1–2 week look-ahead planning window used by facility management. This experiment specifically tests the preliminary robustness of the heuristic in handling a realistic batch of heterogeneous projects compared to the facility's current static planning approach. The reported execution timelines, cycle times, and throughput values are simulation-derived performance measures obtained by applying the baseline and heuristic-controlled schedules to the calibrated testbed.

The comparative assessment demonstrates a marked enhancement in flow efficiency. The application of heuristic-derived buffers decreased the average Cycle Time (Lead Time) by 37.7%, declining from a baseline of 328.65 minutes to 204.70 minutes. This substantial reduction signifies a considerable drop in Work-in-Process (WIP) accumulation throughout the production floor. Notably, this improvement in flow speed was realized while maintaining the system's production rate; the heuristic-controlled throughput (3.617 panels/hour) was statistically equivalent to the baseline (3.668 panels/hour). Following the fundamental principles of Factory Physics [19], a reduction in Cycle Time (*CT*) coupled with constant Throughput (*TH*) mathematically necessitates a reduction in Work-in-Process (*WIP*) (Little's Law: $WIP = TH \times CT$). In the context of volumetric construction, where floor space is the primary constraint, this reduction in on-line inventory is critical. It implies that the facility achieves the same output with significantly less physical clutter, reducing material handling risks and freeing up capital, thereby validating the heuristic's value beyond simple output metrics. The key simulation outputs considered in this evaluation are project execution timelines, average cycle time, and throughput; these are reported quantitatively in the text and visually compared in Figure 3. The Baseline schedule (Blue) displays extended project durations accompanied by significant temporal overlaps. These overlaps characterize periods of intense congestion where multiple projects contend for constrained buffer capacity, producing the "deadlocks" endemic to the facility's existing operations.

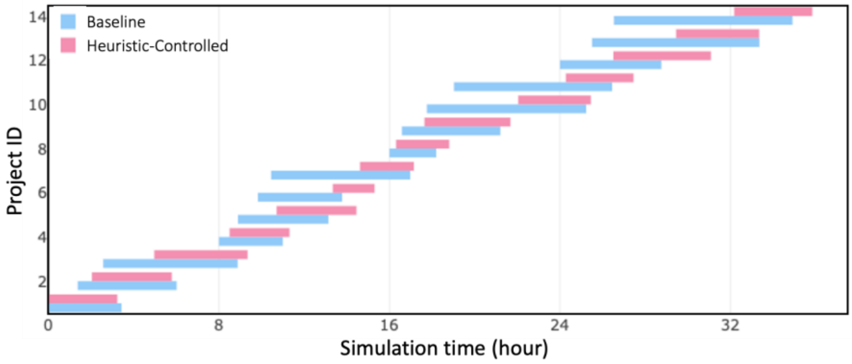

Figure 3: Comparative project execution timeline for the 14-project validation week: Baseline schedule vs. Heuristic-Controlled schedule.

By contrast, the Optimized schedule (Pink) exhibits more compressed execution windows with regular, orderly spacing. By introducing scientifically calibrated lags informed by the bottleneck intensity (M_i), the heuristic efficiently "harmonizes" the flow. Remarkably, the total Makespan of the production week remained essentially unchanged. This counterintuitive finding demonstrates that the "safety buffers" introduced by the heuristic do not postpone overall batch completion. Rather, they redirect unproductive "congestion time" (waiting within the line due to blockage) into productive "waiting time" (waiting outside the line prior to commencement). In volumetric construction, where floor space represents the limiting factor, the 38% improvement in Cycle Time bears strategic significance. By synchronizing production velocity with the stochastic properties of the bottleneck, the framework eliminates the cascading delays that characterize static schedules. The findings establish that disciplined planning can stabilize operations and reduce WIP liability while maintaining the facility's total weekly output capacity.

5 Conclusion

This study addresses the scheduling challenge in High-Mix Low-Volume (HMLV) modular construction, where product variability frequently disrupts static plans. Using a Design-Science Research (DSR) approach, it develops and evaluates a Surrogate-Based Scheduling Framework that translates complex simulation-optimization results into a practical daily planning rule. The framework follows a structured design cycle. A Hierarchical Stochastic Demand Model generates diverse synthetic production scenarios reflecting the variability of Engineer-to-Order projects. These scenarios are evaluated in a calibrated simulation testbed to identify the optimal inter-project buffers required to absorb system variability. The resulting decisions are then converted into a simplified surrogate-based heuristic, enabling planners to size safety margins based on project workload intensity without real-time optimization. Evaluation on a real wood-frame wall panel assembly line showed that the derived "Peak Workload" heuristic reduced average Cycle Time by 37.7% while maintaining statistically equivalent throughput. Under Little's Law, this implies a substantial reduction in Work-in-Process (WIP), which is especially valuable in space-constrained prefabrication environments where congestion limits operational efficiency. The study also highlights several limitations. The current surrogate uses a linear regression model ($R^2 \approx 0.657$) to preserve interpretability under limited planning-stage information, but future work should explore richer features and non-linear models to better capture buffer-sizing behavior. In addition, validation was limited to a one-week look-ahead horizon, so longer-term testing is needed to confirm robustness and generalizability. Finally, the optimization used fixed objective weights; future research should support dynamic weight adjustment so managers can adapt scheduling priorities to changing operational conditions.

6 Acknowledgement

This research is supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) Alliance grant ALLRP 581074–22, Discovery Grant RGPIN-2020-04420, and NSERC Alliance—Alberta Innovates Advance Program 232403286. All findings and conclusions expressed in this paper are those of the authors and do not reflect those of the contributors.

7 References

[1] Abdul Nabi M, El-adaway IH. Modular construction: Determining decision-making factors and future research needs. Journal of Management in Engineering 2020;36:04020085.

[2] Gosling J, Naim MM. Engineer-to-order supply chain management: A literature review and research agenda. Int J Prod Econ 2009;122:741–54.

[3] Salama T, Salah A, Moselhi O. INTEGRATED SCHEDULING OF MODULAR CONSTRUCTION USING "MOD-SCHEDULER." Proceedings of the 7th International Canadian Society for Civil Engineering Construction Conference (CSCE 2019), Montreal, Canada, 2019.

[4] Pourrahimian E, Salhab D, Fayek AR, Hamzeh F. Fuzzy-Chaos Framework for Analyzing and Quantifying Uncertainty in Labour Productivity in Complex Construction Environments. ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction, vol. 42, IAARC Publications; 2025, p. 571–9.

[5] Salama T, Salah A, Moselhi O. Alternative scheduling and planning processes for hybrid offsite construction. ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction, vol. 33, IAARC Publications; 2016, p. 1.

[6] Eltahan A, Lee G, Hamzeh F. Exploring Cognitive Load and Task Complexity in Dynamic Tracking Tasks: Insights for Construction Workflows. ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction, vol. 42, IAARC Publications; 2025, p. 532–9.

[7] Ballard G, Tommelein I. Current process benchmark for the last planner system. Lean Construction Journal 2016;89:57–89.

[8] She B, Chen B, Hall NG. Buffer sizing in critical chain project management by network decomposition. Omega (Westport) 2021;102:102382.

[9] Ma G, Wang A, Li N, Gu L, Ai Q. Improved critical chain project management framework for scheduling construction projects. J Constr Eng Manag 2014;140:04014055.

[10] Moghimi N, Arshang O, Hamzeh F. Digital twins in offsite construction: Current implementations, challenges, and future pathways. Autom Constr 2026;183:106820.

[11] Wu C, Jiang R, Li X. Integration of BIM and computer simulations in modular construction, A case study. Modular and Offsite Construction (MOC) Summit Proceedings 2016.

[12] Matsumoto K, Yamaguchi A, Oka T, Yasumoto M, Hara S, Iida M, et al. Simulation-based reinforcement learning approach towards construction machine automation. ISARC. Proceedings of the International Symposium on Automation and Robotics in Construction, vol. 37, IAARC Publications; 2020, p. 457–64.

[13] Araújo GR, Gomes R, Gomes MG, Guedes MC, Ferrão P. Surrogate models for efficient multi-objective optimization of building performance. Energies (Basel) 2023;16:4030.

[14] Zhang L, Plathottam S, Reyna J, Merket N, Sayers K, Yang X, et al. High-resolution hourly surrogate modeling framework for physics-based large-scale building stock modeling. Sustain Cities Soc 2021;75:103292.

[15] Soman RK, Molina-Solana M, Whyte JK. Linked-Data based Constraint-Checking (LDCC) to support look-ahead planning in construction. Autom Constr 2020;120:103369.

[16] Hevner AR, March ST, Park J, Ram S. Design science in information systems research. MIS Quarterly 2004:75–105.

[17] Moghimi N, Kashani H. Probabilistic Modeling to Develop the Optimal Seismic Resilience Enhancement Strategy for Industrial Facilities. Journal of Infrastructure Systems 2026;32:04025041. https://doi.org/10.1061/JITSE4.ISENG-2754.

[18] Moghimi N, Kashani H. A probabilistic framework to design the financing scheme for resilience enhancement of infrastructure systems. Reliab Eng Syst Saf 2025;262:111179.

[19] Hopp WJ, Spearman ML. Factory physics. Waveland Press; 2011.