Transforming Construction with Off-site Methods and Technologies - TCOT 2026 Conference

August 18–20, 2026, Newcastle upon Tyne, UK

MULTI-OBJECTIVE WORKFORCE ALLOCATION FOR OFF-SITE CONSTRUCTION ASSEMBLY USING PARALLEL SIMULATED ANNEALING

Sbaragli, A.[1,*], Xie, X.[2], and Kassem, M[2]
[1]School of Architecture, Technology and Engineering, University of Brighton, Brighton, BN2 4AT, UK
[2]Department of Engineering, Newcastle University, Newcastle, NE1 7RU, UK
[*]a.sbaragli@brighton.ac.uk

Abstract: This paper proposes a scalable optimisation framework for workforce allocation in off-site construction assembly. The method supports daily staffing decisions in a job shop environment where steel-frame products are assigned to parallel stations. Economic performance and delivery reliability are captured by labour cost, total tardiness, and delay severity. The framework combines a production model, parallel Pareto Simulated Annealing, and context-aware neighbourhood moves. These searches adjust the staffing plan by redistributing workers, opening additional stations, or closing underutilised capacity based on tardiness pressure, slack and utilisation. The method is validated on an off-site construction case study with 1769 structural steel frames, heterogeneous assembly workloads, and seven delivery deadlines. Results show that redistribution moves are ineffective in reducing delays against planned due dates. Configurations with station opening/closing eliminate tardiness and delay severity at low labour cost, converging in about 2 hours.

Keywords: production scheduling; off-site construction; human-centric; simulated annealing

1 INTRODUCTION

Off-site construction (OSC) is increasingly recognised as a strategic pathway to improve operations by relocating value-added activities to controlled factory settings, where prodocution outputs are assembled or fitted before delivery and site installation (Hadi et al. 2023; Zohourian et al. 2025). This shift enables more repeatable and controllable production when OSC is planned as an integrated system, with Design for Manufacture (DfMA), Building Information Models(BIM), and lean principles linking design, information flow, and waste reduction (Hadi et al. 2023; Waheed et al. 2025; Mujahed et al. 2024). However, high project-based variabilities require deeper levels of production control (Peiris et al. 2023; Aghajamali et al. 2026). These operating conditions make digital production control essential for improving visibility of manual work, resource utilisation, and process deviations (Hadi et al. 2023; Sbaragli and Nardello 2025; Mujahed et al. 2024). Within this agenda, production scheduling converts project demand and product-level work content into executable factory decisions. The definition of job sequences, workstation assignments, resource levels, and timing enables OSC job shops to improve makespan, labour utilisation, delivery reliability, and cost performance while reducing disruptions to site assembly (Peiris et al. 2023; Aghajamali et al. 2026; Kim et al. 2020; Tao et al. 2026). However, workforce allocation remains under-explored as a key decision in labour-intensive OSC assembly, despite its direct effect on capacity, delivery performance, and Construction 5.0's human-centric emphasis (Hadi et al. 2023; Sbaragli and Nardello 2025). To address this gap, this paper proposes a multi-objective workforce allocation framework that combines parallel Pareto Simulated Annealing (SA) with context-aware neighbourhood moves. The framework generates day-by-day and station-by-station staffing plans that balance labour cost, total tardiness, and delay severity. The remainder of this work is organised as follows. Section 2 reviews OSC scheduling and optimisation studies. Section 3 formulates the workforce allocation problem. Section 4 presents the proposed parallel SA framework. Section 5 discusses the case study and benchmark results. Section 6 concludes the paper and outlines future research opportunities.

2 LITERATURE REVIEW

OSC production scheduling is an increasingly investigated area, with applications ranging from job sequencing to resource allocation. Existing studies use methods such as discrete-event simulation (DES) and metaheuristics (e.g., SA and genetic algorithms) to optimise factory operations (Peiris et al. 2023).

A first stream of studies addresses due-date reliability and dynamic production control. Kim et al. (2020) propose dynamic scheduling model for precast concrete production to minimise tardiness. The developed DES model benchmarks different dispatching rules, where Earliest Due date considering Level of Uncertainty emerges as the most performing candidate demonstrating tardiness reductions between 65 and 129.5 hours. The methodology unlocks scheduling flexibility by adapting production priorities to uncertain site-driven due dates. However, the model does not explicitly capture the strategic role of human operators in the production system. A similar limitation is observed in Bhatia et al. (2023), who minimise makespan in a wood-based wall-panel production line. The methodology feeds workstations' processing time into a genetic algorithm (GA) with 100 generations, a population of 30, a crossover rate of 0.8, and a mutation rate of 0.2. The most performing schedule shrinks the makespan from 4359 to 4352 minutes, converging after approximately 90 generations. However, the improvement is marginal, with only a 7-minute makespan reduction, and the validation remains limited to a small dataset of 46 wall panels. In addition, no computation time is reported, which limits the assessment of the model's deployability in OSC environments.

A second optimisation stream focuses on resource utilisation and cycle times. For instance, Hou et al. (2025) proposes a hybrid metaheuristic to improve modular design and production scheduling for the housing sector. The method combines GA global exploration with SA local refinement to improve module cohesion, reduce inter-module coupling, and optimise machine allocation. The hybrid model reduces production cycles by up to 30% across datasets of up to 100 instances, while improving resource utilisation and cost control. Despite notable operational performance improvements, the study remains limited by relatively small experimental instances, which may not reflect the scale and variability of project-based OSC production. While cost and resource optimisation are machine-driven, human operators are largely absent as explicit workforce decision variables.

A final stream explicitly incorporates human resources into OSC production planning, reflecting the labour-intensive nature of prefabricated assembly. For example, Aghajamali et al. (2026) develop a BIM and numerical control framework to estimate labour hours and optimise construction-order sequencing in structural steel prefabrication. The dataset includes around 7000 steel frames, mapped against production features such as assembly and weld length. The framework combines statistical labour-hour prediction, DES-based production simulation, and a similarity-based heuristic to reduce repeated operations and improve production flow. The model reduces labour cost by 12.99% compared with the reference planning approach. However, workforce is mainly treated as an estimated workload for sequencing, rather than as a strategic decision variable to meet site installation due dates. A similar limitation can be appreciated in Xiong and Peng (2026), that propose a BIM-driven improved particle swarm optimisation (PSO) model to optimise resource allocation in prefabricated building construction. BIM is used to extract multidimensional resource-demand data, while the improved multi-vector PSO expands the particle search space to enhance global optimisation and avoid premature convergence. The model is validated on a small prefabricated building case with five tasks and two resources. Results show that the maximum resource daily demand decreases from 8 to 6 and from 5 to 4 demand units/day, while labour demand is reduced by up to 15 workers on a single day. More explicitly, Tao et al. (2026) address part of this limitation by modelling workforce skills and availability as active constraints in precast component production scheduling. The authors combine an integer linear programming model with a Deep Q-Network algorithm to model outsourcing decisions, production-line scheduling, and multiskilled workforce assignment. The algorithm is validated using real data from a precast manufacturer, considering both off and high-peak demand scenarios. Results show average monthly savings above 31000 CNY compared with empirical planning, demonstrating the value of integrating workforce constraints into production decision-making. Human labour is modelled through multiskilled worker availability and task-assignment constraints, allowing the model to decide which qualified workers process internal orders. However, staffing is not treated as a daily station-level capacity decision, where the number of operators assigned to each workstation directly controls processing speed, labour cost, and delivery performance. In summary, the reviewed production scheduling applications provide valuable methodologies to improve the operational efficiency of OSC environments but their applicability can be further enhanced from two key perspectives:

- **Workforce modelling:** labour is often represented indirectly as estimated man-hours, resource demand, or skill-assignment constraints, rather than as a strategic staffing variable that controls workstation capacity, labour cost, and delivery performance.
- **Scalability and deployability:** several studies rely on small or medium-sized instances, report limited computation-time evidence, or focus on project-level resource planning rather than station-level and job shop control.

Building on this discussion, this paper proposes a scalable, human-centric workforce allocation framework for OSC assembly. The method combines parallel Pareto SA with context-aware neighbourhood moves that redistribute workers and open or close workstation capacity according to schedule pressure, slack, and utilisation. The framework is validated on an industrial dataset of 1769 structural steel frames, balancing labour cost, total tardiness, and delay severity. Benchmark results show that the proposed move families achieve zero tardiness and zero delay severity while converging in less than 2.5 hours.

3 PROBLEM DESCRIPTION

This work addresses a workforce allocation and delay management problem in OSC factories. The production system is modelled as a job shop environment in which final products are assembled from kits of parts across multiple stations (see Fig. 1). Assembly workstations may operate in parallel, and each may be staffed by a variable number of operators. When several operators are assigned to the same workstation, they work jointly on the same assembly task, which increases the effective processing rate and reduces the total assembly time. Staffing plan allocations specify the number of operators assigned to each workstation with daily granularity. Each product is associated with an assembly deadline, representing the latest acceptable completion time before shipment to the construction site. Delays occur whenever a product is completed after its deadline, potentially affecting subsequent operations and overall project performance. To evaluate production schedule performance, two complementary time-driven indicators are considered. While the tardiness defines the delay beyond the assigned deadline, the severity expresses the mean tardiness across delayed jobs only. These indicators capture different dimensions of delivery disruption, since the operational consequences of delays depend on both their magnitude and distribution across products. For instance, decision-makers may prefer a schedule in which a small number of products experience limited delays rather than one in which a larger proportion of products are delayed for several days. Besides delivery metrics, labour cost captures the job shop's economic impact on projects' cash flows. Therefore, this multi-objective problem is formulated to determine staffing plans that minimise operational delays while limiting labour expenditure.

3.1 Parameters

The assembly model is defined by a set of indices and parameters describing the production system, the assembly jobs, and the operational constraints governing workforce allocation. Let $j = \{1, ..., J\}$ denote the set of assembly jobs to be processed within the planning horizon. Each job corresponds to a kit of structural components requiring manual assembly operations. These activities are performed in one of the

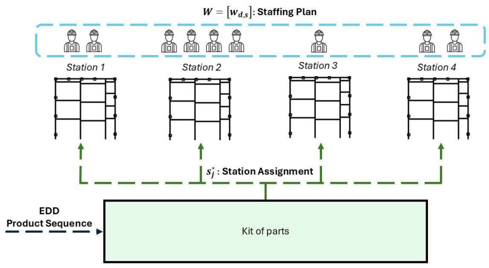


Figure 1: Assembly off-site construction job shop

available workstations, denoted by $s = \{1, ..., S\}$. The workstations operate independently, while sharing the available workforce resources. Workforce allocation decisions are defined over a planning horizon discretised into working days, represented by the set $d = \{1, ..., D\}$. Each job $j \in J$ is characterised by a specific assembly workload determined by the number of screws (e.g., q_j)required to assemble the corresponding steel frame panels. Each job is also associated with a delivery deadline δ_j, representing the latest acceptable completion time. Meeting these deadlines is crucial to preserving the continuity of downstream construction activities (e.g., site installation) and avoiding potential penalties that erode project margins. The assembly process is governed by several operational and time-driven parameters. The parameter b denotes the average time required by one operator to install one screw under standard working conditions. Each operator can work at most $T^{\max}$ hours per day, while the daily productive capacity of a workstation depends on the number of assigned operators. Staffing is further constrained by $w^{\max}$, which defines the maximum number of operators that can be simultaneously assigned to a single workstation. The workforce efficiency exponent α captures the non-linear productivity gain from assigning multiple workers to the same workstation, accounting for coordination requirements and workspace limitations. Labour cost is represented by the hourly cost parameter c^L, which reflects the economic impact of workforce utilisation.

3.2 Decision Variables

While jobs are sequenced a priori according to the Earliest Due Date (EDD) rule and thus constitute a fixed input to the optimization, the staffing plan $W \in Z^{D\times S}$ represent the core decision variable of the problem. Each entry $w_{d,s}$ denotes the number of operators assigned to $s - th$ station and $d - th$ day. The variable $w_{d,s}$ directly influences the effective processing speed and daily capacity of each workstation. When $w_{d,s} = 0$, that station is not operating. Although job-to-station assignments are treated as fixed within a given solution, changes in station availability may trigger a recomputation of these assignments (see Section 4).

3.3 Multi-objective Optimization

Given the fixed EDD job sequence and a candidate staffing plan, the schedule is evaluated through dynamic dispatch. For each job, the optimizer selects the station that yields the earliest feasible completion time under the current staffing configuration, station availability, and working-day calendar. The developed multi-objective optimisation aims at minimising labour expenditure, total tardiness, and delay severity. The first objective measures the total **labour cost** associated with the workforce allocation plan. Let d^* denote the last calendar day on which at least one job is processed. Since labour cost is incurred only over the effective staffed horizon and on working days, the total labour cost is computed as follows.

$$f_c = c^L T^{\max} \sum_{d \in \mathrm{D}^{\mathrm{wd}}(d^*)} \sum_{s \in \mathrm{S}} w_{d,s} \quad \text{(Eq. 1)}$$

where $\mathrm{D}^{\mathrm{wd}}(d^*) \subseteq \mathrm{D}$ denotes the subset of working days up to and including the last used day d^*. The cumulative **tardiness** represents the second problem objective. Let C_j denote the completion time of job j

and job tardiness is defined as $T_j = \max(0, C_j - \delta_j)$. Therefore the cumulative tardiness is computed as it follows.

$$f_t = \sum_{j\in \mathrm{J}} T_j \quad \text{(Eq. 2)}$$

In addition to cumulative tardiness, the model evaluates delay intensity through a **severity** objective, which distinguishes schedules with similar total tardiness but different delay distribution. Let $\mathrm{J}^+ = \{j \in \mathrm{J} : T_j > 0\}$ denote the set of delayed jobs. The severity objective is computed as follows.

$$f_s = \frac{1}{|\mathrm{J}^+|} \sum_{j\in \mathrm{J}^+} T_j \quad \text{(Eq. 3)}$$

Finally, the optimisation problem is subject to a set of **operational constraints** reflecting the physical limitations of the OSC assembly system.

$$w_{d,s} \in \{0, 1, \dots, w^{\max}\}, \quad \forall d \in \mathrm{D},\ \forall s \in \mathrm{S}. \quad \text{(Eq. 4)}$$

$$v_{d,s} = w_{d,s}^{\alpha}, \quad \forall d \in \mathrm{D},\ \forall s \in \mathrm{S}, \quad 0 < \alpha \leq 1. \quad \text{(Eq. 5)}$$

$$\sum_{d\in \mathrm{D}} \sum_{s\in \mathrm{S}} x_{j,d,s} = P_j, \quad \forall j \in \mathrm{J}. \quad \text{(Eq. 6)}$$

$$\sum_{j\in \mathrm{J}} x_{j,d,s} \leq T^{\max} v_{d,s}, \quad \forall d \in \mathrm{D},\ \forall s \in \mathrm{S}. \quad \text{(Eq. 7)}$$

$$s_j^* = \arg\min_{s\in \mathrm{S}} \hat{C}_{j,s}(W, A_s), \quad \forall j \in \mathrm{J}. \quad \text{(Eq. 8)}$$

$$\sum_{s\in \mathrm{S}} y_{j,s} = 1, \quad \forall j \in \mathrm{J}, \qquad y_{j,s} \in \{0, 1\}, \quad \forall j \in \mathrm{J},\ \forall s \in \mathrm{S}. \quad \text{(Eq. 9)}$$

$$y_{i,s} = y_{j,s} = 1 \Rightarrow B_j \geq C_i, \quad \forall i, j \in \mathrm{J},\ i \prec j,\ \forall s \in \mathrm{S}. \quad \text{(Eq. 10)}$$

Eq. 4 enforces bounded and integer staffing decisions at each station on each working day. Eq. 5 captures the effect of collaborative staffing on workstation productivity through a concave efficiency function. Eq. 6 ensures that the full workload of each job is completed, while Eq. 7 imposes the daily processing capacity of each station. The dispatching rule is defined by earliest feasible completion in Eq. 8. Eq. 9 enforces a unique binary station assignment for each job, with the selected station determined by the dispatching rule. Finally, Eq. 10 enforces non-overlapping processing for jobs assigned to the same station according to the fixed EDD sequence.

4 METHODS

To define optimized $w_{d,s}$ subject to Eq.1-10, the framework integrates a Pareto search with context aware neighbourhood moves to ensure the optimisation remains computationally efficient and grounded OSC operations. The tree step based optimisation workflow is discussed in the following Sections.

4.1 Initialisation

A fixed job sequence is obtained by sorting assembly products with an EDD logic, providing a realistic starting solution for tardiness-oriented scheduling problems. Then, the staffing plan matrix $w_{d,s}$ is initialised by activating two stations and assigning two workers each until the job sequence is fully assembled. During initialisation, jobs are assigned to workstations by allocating each job to the next available assembly station. The initial temperature is calibrated from random neighbours of the initial solution. For each neighbour, objective changes are normalised, and the worsening components are combined into a single deterioration score. The median positive deterioration is then used to set T_0, so that a typical worsening move is accepted with probability p_0. The remainder of this section presents the context-aware neighbourhood design and the parallel Pareto SA algorithm, respectively (see Sections 4.2 and 4.3).

4.2 Neighbour Moves

The neighbourhood operator perturbs the staffing matrix by probabilistically selecting one of three move families, namely **redistribution**, **opening** or **closing**, and then sampling a specific move within the selected family. Move selection is guided by the current solution context, which captures day-level tardiness pressure, day-level slack, station-day utilisation, and the last effective production day. Day-level slack measures the positive time margin remaining before a job's deadline, defined as $\max(0, \delta_j - C_j)$ and distributed across its processing days. Tardiness pressure identifies critical periods and is propagated to preceding days, allowing capacity changes to be introduced before delays materialise. Candidate moves are then restricted to working days within a focused horizon around the active production window. The first neighbour move family is **redistribution** that reallocates labour capacity within the currently active assembly stations. These moves are designed to rebalance labour where schedule pressure is highest. Worker additions are biased toward high-pressure days, including nearby upstream days, whereas worker removals are biased toward low-pressure, high-slack periods with comparatively low utilisation. This family is designed with the following operators.

- **Rebalance Day Active:** this operator is selected 38% of the time and transfers one worker between two active stations on the same day to reallocate intra-day capacity without changing total daily staffing.
- **Shift Same Station:** This move accounts for 30% of redistribution selections and shifts one worker for the same station across two days to move effort toward higher-pressure periods.
- **Single Station Active:** increases or decreases by one the staffing level of a single active station-day, with the direction and location biased by pressure/slack and utilisation signals. The selection probability of this move is equal to 22%
- **Station Block Active:** Representing 10% of redistribution selections, increases or decreases by one the staffing level over a short consecutive block (e.g., 2 to 3 days) for an already active station, enabling medium-scale temporal rebalancing.

Opening moves represent the second neighbour family, increasing available capacity by activating a previously unused workstation on one or more production days. While candidate days are selected from high-pressure periods with rollback, the staffing level assigned to an opened station is sampled in $[1, w^{\max}]$ with a bias toward larger values under stronger pressure. When selected, this family samples the following operators.

- **Open Station Day:** 35% of selected moves activate a previously inactive workstation on one critical day and assigns a staffing level, bounded by the constraint in Eq. 4 .
- **Open Station Block:** activates a previously inactive workstation over a short consecutive block (e.g., 2 to 3 days), using the same staffing level across the block. This move is selected with a probability of 65%.

Finally,**closing** moves reduce capacity by deactivating an active workstation on a single day or short day block. They are applied only in low-risk periods with positive slack, negligible pressure, and low utilisation, while feasibility checks ensure that at least one station remains active on each affected day.

- **Close Station Day:** deactivates an active workstation on a specific day by setting its staffing level to zero. This move is selected with a probability of 85%.
- **Close Station Block:** Used in the remaining 15% of closing selections, this operator deactivates an active

workstation over a short consecutive block (e.g., 2 to 3 days) by setting its staffing level to zero throughout the block.

Job-to-workstation assignments are recomputed only when opening and closing moves modify the set of active workstations. Redistribution moves keep the station structure unchanged and only schedule timing is re-evaluated (e.g., job tardiness). This design reflects a deliberate computational trade-off aimed at reducing optimisation run time, accepting that assembly imbalances are addressed through redistribution moves.

4.3 Parallel Simulated Annealing

The optimisation procedure is structured into M outer temperature iterations, each comprising the evaluation of N candidate neighbours. Parallelisation is achieved using a process pool of P workers or threads, exploited for both neighbour generation and schedule evaluation. To mitigate overhead, computations are performed in batches of size equal to $B = P \cdot chunk$. A chunk denotes a group of candidate solutions generated and evaluated in parallel before synchronising with the main search process. For each batch, a move context is generated from the current solution and broadcast to all parallel workers, enabling informed neighbour generation. Each worker independently samples a candidate staffing plan using the previously discussed neighbour moves. The resulting solution is then evaluated against the multi-objective function, together with the realised job-to-station assignments. Infeasible solutions, such as an insufficient staffing horizon, are highly penalised and thus excluded from the acceptance phase. The acceptance mechanism follows a Pareto-based approach. If a candidate solution strictly improves the current solution, it is accepted deterministically. Otherwise, if the candidate is dominated by the current solution, acceptance follows a temperature-dependent probability based on its normalised deterioration, so larger worsening moves become less likely as the search cools. When the candidate and current solution are mutually non-dominated, the candidate is accepted as a valid trade-off. In cases where multiple candidates within the same batch satisfy the acceptance criteria, a single solution is selected. The selection approach follows a hierarchical rule, where preference is given first to dominating solutions, then to non-dominated trade-offs, and finally to probabilistically accepted dominated solutions. Ties within each category are resolved lexicographically, prioritising lower tardiness, followed by cost, and then severity. An external Pareto archive is maintained throughout the search process to store all non-dominated solutions encountered. A candidate is inserted into the archive only if it is unique and not dominated by any existing member. Upon insertion, any archive entries that are dominated by the new solution are removed, ensuring that the archive remains a consistent approximation of the Pareto front. At the end of each outer iteration, the temperature is reduced according to a geometric cooling schedule. Early termination of the algorithm is governed by a sequence of plateau-based criteria reflecting the hierarchical importance of the objectives. The procedure first monitors improvements in tardiness. Only when tardiness has stagnated for a predefined number of iterations, the optimiser evaluates improvements in cost. Similarly, once cost improvements plateau, attention shifts to severity. The search terminates when no further improvement is observed across this hierarchy for the prescribed patience thresholds, indicating convergence with respect to the lexicographic objective structure.

5 RESULTS

This section validates methodology in a UK-based OSC job shop and benchmarks three parallel SA configurations across key operational metrics (see Sections 5.1 and 5.2).

5.1 Case Study

The parallel SA is tested and validated on a dataset of 1769 structural steel frames with starting production day 5/05. The core challenge for the OSC company is to dimension the workforce daily by minimising product delays against planned deadlines while containing labour costs. The job shop assembly station operates in parallel, each staffed with up to four skilled operators. Based on assembly drawings information, operators perform manual stud fitting and screwing tasks, along with any necessary rework to correct modeling errors. Completed frames are categorized into 16 product types and follow distinct production routes. While product

Category	Parameter	Value	Unit
PM	q_j	$\in [2, 204]$	screws/frame
	δ_j	$\in [12/05, 2/07]$	Date
	b	60	sec/(screw$\times$op)
	$T^{\max}$	8	h/day
	$w^{\max}$	4	op/station
	S	4	stations
	α	0.9	-
	c^L	20	£/h
	$start_{day}$	5/05	Date
SA	p_0	80	%
	M	50	iteration
	N	17690	iteration
	α_{cool}	0.98	-
WS	P	12	threads
	$chunk$	30	-

Table 1: Case study parameters

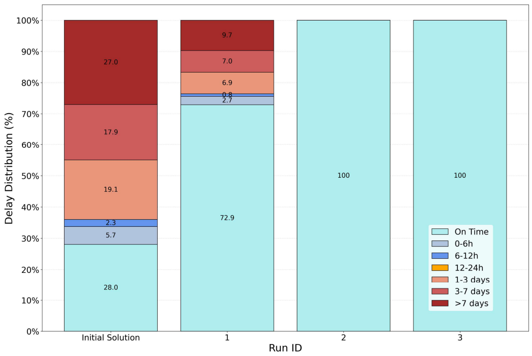
Run ID	Redistribute	Opening	Closing	Tardiness (h)	Cost (£)	Severity (h)	Run Time (h)
Initial	-	-	-	176955.22	30720	139	-
R1	100	0	0	50836.57	31200	106.13	2.01
R2	80	20	0	0	32160	0	1.6
R3	**70**	**15**	**15**	**0**	**31200**	**0**	**2.18**

Table 2: Parallel Simulated Annealing configurations performance benchmark

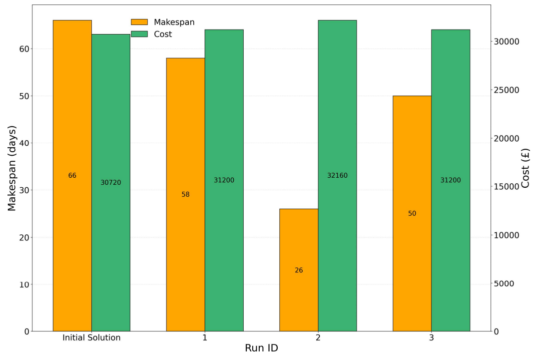
IDs 1 and 2 are dispatched directly to construction sites for installation, the remaining products represent subsystems for other in-plant job shops. Tab. 1 lists the problem's parameters grouped into three categories: production model (PM), parallel SA, and the computing workstation configuration (WS). The WS workstation is equipped with an AMD Ryzen 5 3600 six-core processor with two threads per core and 16 GB of RAM.

5.2 Multi-objective benchmarking of workers allocation

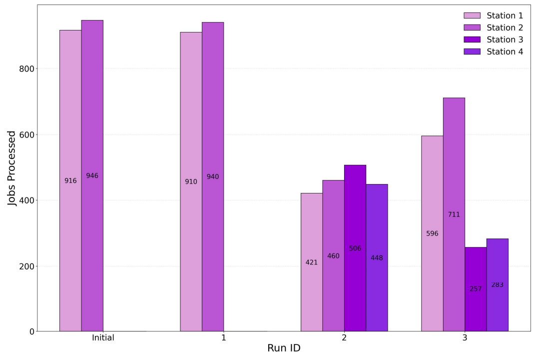
This section presents a comparative analysis of three parallel SA configurations under different neighbourhood-family probability settings (see Tab. 2). The benchmark is initialised with a greedy solution that sorts structural frames by EDD and equally splits four workers across two assembly stations. This initial plan yields the lowest direct labour cost of 30720 £, but exhibits poor service performance, with 176955.22 h of total tardiness and a delay severity of 139 h.This understaffed solution is improved by $R1$ that activates **redistribution** moves only. Although time-driven objectives improve, both tardiness and delay severity remain far from optimal, demonstrating that increasing only the number of workers in the original staffing plan is insufficient. **Opening** moves are required to provision new stations around production bottlenecks, while **closing** ones are incorporated to contain labour costs by accepting larger makespans only when all deadlines are satisfied. $R2$ and $R3$ provide optimal delivery service levels, achieving zero tardiness and severity. The core distinction between these two configurations lies in the cost and run-time trade-off. $R2$ converges in 1.6 h by reducing the proportion of moves triggering frame-to-workstation assignment computation from 30% to 20%, whereas $R3$ achieves a marginally lower labour cost of 31200 £over the same time window. This multi-objective discussion is further enriched by evaluating several production KPIs for each run. Fig. 2a



(a) Delay distribution



(b) Makespan and Labour Cost



(c) Stations production output

Figure 2: Production KPIs across Parallel Simulated Annealing Run IDs

demonstrates how the proposed family-oriented neighbourhood moves enable the optimiser to eliminate production delays. Indeed, $R2$ and $R3$ achieve full on-time delivery from an initial state in which more than 60% of products were completed at least one calendar day late. Fig. 2b reveals that comparable labour expenditures propagate differently onto production makespan. Although $R2$ achieves the shorter makespan of 26 calendar days, production stakeholders prefer $R3$ due to its lower operational cost and greater flexibility in allocating the remaining workforce across building sites, thereby reducing reliance on temporary external recruitment. Both $R2$ and $R3$ maintain the same number of active workers and workstations during the bottleneck period spanning 09/05 to 16/05. Beyond this critical week, however, $R2$ leverages at least three active stations with daily workers ranging from 5 to 16. This dynamic staffing allocation directly governs the throughput of each assembly station (see Fig.2c). Starting from more than 900 frames processed per station in the initial solution and in $R1$, the latter two configurations systematically introduce a third and fourth assembly resources to reduce both tardiness and delay severity to zero, while maintaining a resource utilisation rate no lower than 98.7%. To conclude, $R3$ is selected as the optimal solution, as it minimises staffing costs over the modelled production period and generates a 50-day production schedule within approximately 2 h of computational time, satisfying operational planning requirements.

6 CONCLUSIONS & FURTHER RESEARCH

This paper proposes a multi-objective optimisation framework to improve workforce allocation and operational reliability in OSC assembly, modelling the production system as a job-shop environment where daily staffing plans control workstation capacity, labour expenditure, and delivery performance. The proposed parallel Pareto SA algorithm combines context-aware neighbourhood moves that redistribute workers, open additional stations, and close underutilised capacity according to tardiness pressure, slack, and utilisation.

Starting from an initial solution with more than 70% of steel frames delivered late, the best-performing SA eliminates all operational delays at an increased cost of 480 £. The framework provides production managers with actionable day-by-day staffing plans, supporting more informed decisions on capacity, worker release, and deadline protection without excessive labour expenditure. Future research should incorporate stochastic task execution times to capture operator variability, learning effects, and rework, while integrating the framework into the company's software ecosystem to automate data collection and support on-demand workforce planning. Finally, human-centric parameters such as fatigue should be included in the model to safeguard the physical resilience of workers and fully align with Construction 5.0 principles.

ACKNOWLEDGEMENTS

The authors thank Paolo Rech (University of Trento) for the support in providing the computing infrastructure.

REFERENCES

Aghajamali, Kamyab, Alireza Rahimi, Saeid Metvaei, Zhen Lei, Qian Chen, Alaeldin Suliman, and Ali Pirzad. 2026. "Production planning with an enriched BIM-Based heuristic construction order model to improve the efficiency of complex prefabricated projects." *Journal of Construction Engineering and Management* 152 (6): 04026057.

Bhatia, Angat Pal Singh, Osama Moselhi, and SangHyeok Han. 2023. "Optimized Production Scheduling for Modular Construction Manufacturing." In *Proceedings of the 40th International Symposium on Automation and Robotics in Construction (ISARC 2023),* 270–277. International Association for Automation / Robotics in Construction.

Hadi, Ali, Franco Cheung, Solomon Adjei, and Anmar Dulaimi. 2023. "Evaluation of Lean Off-Site Construction Literature through the Lens of Industry 4.0 and 5.0." *Journal of Construction Engineering and Management* 149 (12): 03123007.

Hou, Xiaodong, Weidong Ji, Weicheng Ji, Zhengyi Zhu, and Xueli Lv. 2025. "Modular Design and Production Scheduling Optimization of Tourism Equipment Housing Based on Simulated Annealing Algorithm." In *The 4th International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS 2025),* 41–46. Institution of Engineering / Technology.

Kim, Taehoon, Yong-Woo Kim, and Hunhee Cho. 2020. "Dynamic Production Scheduling Model under Due Date Uncertainty in Precast Concrete Construction." *Journal of Cleaner Production* 257:120527.

Mujahed, Layla, Gang Feng, and Jianghua Wang. 2024. "Transforming Construction: Integrating Off-Site Techniques and Advanced Technologies." *International Journal of Civil and Architectural Engineering* 18 (9): 368–377.

Peiris, Achini, Felix Kin Peng Hui, Colin Duffield, and Tuan Ngo. 2023. "Production Scheduling in Modular Construction: Metaheuristics and Future Directions." *Automation in Construction* 150:104851.

Sbaragli, Andrea, and Matteo Nardello. 2025. "From Sensing to Segmentation: Transformer-Based Worker Activity Recognition for Industrial Assembly." *IEEE Access* 13:207475–207487.

Tao, Sha, Junying Yu, Jianbo Zhu, Shiwei Chen, and Xiao Li. 2026. "Precast Component Production Scheduling Considering Outsourcing and Multiskilled Resources." *Journal of Construction Engineering and Management* 152 (5): 04026040.

Waheed, Muhammad, Mutahir Abbas, and Aftab Hameed Memon. 2025. "Just-In-Time Importance in the Construction Industry." *AMPLITUDO: Journal of Science and Technology Innovation* 4 (2): 99–102.

Xiong, Bei, and Chuang Peng. 2026. "Optimization of Construction Resources for Prefabricated Buildings Based on an Improved PSO Algorithm Driven by BIM." In *Civil Engineering and Smart Structure Technology: Proceedings of the 2nd International Symposium (CESST 2025), Zhengzhou, China, 5–7 December 2025,* edited by L. Xie, S. Wang, and S. M. Anas, 86:437–446. Advances in Transdisciplinary Engineering. IOS Press.

Zohourian, Mohammadamin, Apurva Pamidimukkala, Sharareh Kermanshachi, and Deema Almaskati. 2025. "Modular Construction: A Comprehensive Review." *Buildings* 15 (12): 2020.