Appendix J: Hyperparameter Selection for the DHPP Model

J.1. Parameter Ψ Calibration for β^i Estimation in DHPP Model

The parameter Ψ is a hyperparameter in the DHPP model, selected via grid search using 10-fold cross-validation on the training dataset over the range $\{0.05, 0.10, \ldots, 0.90, 0.95\}$. To illustrate, we present results for two representative cases: (i) predicting the picking time for type 1 product during early morning shift, and (ii) predicting the performance inconsistency during the same time periods. The corresponding MAPEs and statistical metrics of the optimal β^i (recalling that we have $n = 202$ pickers, with an optimal β^i determined for the i^{th} picker) are reported in Tables J13 and J14, respectively. Based on these two tables, the optimal value of Ψ is selected to be 0.4, as it yields the lowest MAPEs in both cases. This selection result reflects a moderate level of uncertainty in the data. The moderate values of both $\Psi = 0.4$ and the associated optimal β^i indicate that the picking performance exhibits moderate deviations from the expected values, which aligns well with the stochastic nature of real-world operational environment.

Table J13 MAPEs and β^i of DHPP for Picking Time Prediction with respect to Ψ

Ψ	MAPE	β^i				
		Min	Max	Mean	Median	StdDev
0.05	0.1118	21.9536	45.8542	33.8832	34.1758	4.4087
0.1	0.1108	15.5826	32.3761	24.1757	24.3226	3.0979
0.15	0.1096	12.7715	26.4258	19.8760	19.9882	2.5160
0.2	0.1055	11.0795	22.9078	17.3048	17.3986	2.1731
0.25	0.1034	9.9505	20.5075	15.5482	15.6488	1.9382
0.3	0.1010	9.1227	18.7452	14.2561	14.3543	1.7650
0.35	0.0999	8.4860	17.3692	13.2502	13.3359	1.6284
0.4	**0.0912**	**7.9772**	**16.2579**	**12.4396**	**12.5103**	**1.5178**
0.45	0.1007	7.5578	15.3420	11.7694	11.8299	1.4262
0.5	0.1027	7.2025	14.5796	11.2013	11.2614	1.3485
0.55	0.1039	6.8909	13.9180	10.7062	10.7655	1.2823
0.6	0.1045	6.6158	13.3387	10.2744	10.3300	1.2249
0.65	0.1053	6.3735	12.8278	9.8934	9.9459	1.1742
0.7	0.1060	6.1596	12.3744	9.5549	9.6047	1.1290
0.75	0.1066	5.9705	11.9683	9.2512	9.3006	1.0883
0.8	0.1072	5.7998	11.5968	8.9747	9.0251	1.0516
0.85	0.1078	5.6418	11.2580	8.7220	8.7653	1.0182
0.9	0.1088	5.4994	10.9490	8.4903	8.5340	0.9881
0.95	0.1096	5.3668	10.6643	8.2768	8.3163	0.9602

Table J14 MAPEs and β^i of DHPP for Performance Inconsistency Prediction with respect to Ψ

Ψ	MAPE	β^i				
		Min	Max	Mean	Median	StdDev
0.05	0.0447	0.8349	7.3440	2.7124	2.4183	1.2925
0.1	0.0410	0.6225	5.3838	1.9539	1.7357	0.9290
0.15	0.0405	0.5235	4.4873	1.6133	1.4366	0.7660
0.2	0.0400	0.4621	3.9414	1.4084	1.2569	0.6678
0.25	0.0393	0.4172	3.5638	1.2675	1.1338	0.6004
0.3	0.0384	0.3835	3.2817	1.1630	1.0417	0.5503
0.35	0.0367	0.3580	3.0598	1.0814	0.9675	0.5112
0.4	**0.0360**	**0.3367**	**2.8798**	**1.0155**	**0.9060**	**0.4796**
0.45	0.0380	0.3197	2.7301	0.9607	0.8544	0.4534
0.5	0.0393	0.3029	2.6029	0.9140	0.8115	0.4313
0.55	0.0402	0.2880	2.4929	0.8737	0.7763	0.4123
0.6	0.0402	0.2756	2.3965	0.8385	0.7458	0.3956
0.65	0.0402	0.2651	2.3109	0.8074	0.7189	0.3808
0.7	0.0402	0.2560	2.2341	0.7797	0.6948	0.3676
0.75	0.0404	0.2478	2.1651	0.7547	0.6732	0.3557
0.8	0.0407	0.2407	2.1024	0.7322	0.6537	0.3449
0.85	0.0407	0.2343	2.0450	0.7116	0.6358	0.3350
0.9	0.0406	0.2285	1.9924	0.6928	0.6195	0.3260
0.95	0.0405	0.2229	1.9439	0.6753	0.6044	0.3177

J.2. Selection of the Parameters C_1, C_2 and C_3 for DHPP Model Calibration

For the DHPP model, the key parameters C_1, C_2 and C_3 were selected from the range $\{2^{-4}, 2^{-3}, \ldots, 2^{16}\}$. Recall that these parameters control the trade-offs among the four components of the DHPP objective functions: the regularization term, the expected prediction errors of first and second order, and the error deviations for uncertain points.

To illustrate the role of these parameters in balancing the trade-offs, we take the example of predicting the picking time of type 1 product in early morning shift, for presenting the selected parameters C_1, C_2 and C_3 with the corresponding optimal values of each term in the objective function of the DHPP model in Table J15. As shown, the fourth term, which represents the error deviations for uncertain points, along with the first regularization term, are the two most significant contributors, emphasizing the model's robustness. The second term, corresponding to the expected prediction errors of first order, is the third most influential, highlighting its role in maintaining the prediction accuracy. The third term, though also related to prediction accuracy, plays a relatively smaller role, possibly due to its overlap with the second term.

Table J15 The selected parameters C_1, C_2 and C_3 with the related optimal values of each term in the objective of DHPP for predicting the picking time of Type 1 product in early morning

C_1	C_2	C_3	$\boldsymbol{\omega}^T\overline{\boldsymbol{Q}}\boldsymbol{\omega}$	$\frac{C_1}{n}\sum_{i=1}^{n}\bar{\zeta}^i$	$\frac{C_2}{n^2}\bar{\boldsymbol{\zeta}}^T(n\boldsymbol{I}_n-\boldsymbol{1}_n)\bar{\boldsymbol{\zeta}}$	$C_3\sum_{i=1}^{n}\beta^i$
8	2	0.25	185.006	138.080	0.019	621.979

Appendix K: Investigation of the Generalizability of DHPP Model on UCI Public Benchmark Datasets with Noise

In this Appendix, we evaluate the robustness and generalizability of the proposed DHPP model by testing it on publicly-available datasets with added noise. Additional numerical results of all tested models on datasets from our partner warehouse with added noise, can be assessed upon request from the authors.

To further validate the performance of our DHPP model beyond the specific context of JD.com, we have also conducted additional robustness tests using public benchmark Slump, Autoprice, Machine, MPG, Housing, Traffic, Cargo and Lattice datasets from the University of California at Irvine (UCI) repository, which is widely-used in the machine learning field. Table K1 summarizes the dataset characteristics, where the number of points corresponds to the number of pickers in our dataset in the JD.com case. Hence, these datasets allow us to assess the generalizability of our model and ensure its effectiveness in diverse scenarios with much larger datasets. In particular, to account for the uncertainty in the benchmark datasets, we incorporated noise pollution related to each feature of each dataset to generate uncertain datasets. This approach allows us to simulate the effects of uncertainty on the model's performance and test its robustness under conditions similar to those in the JD.com case.

Following the approach outlined in Balasundaram and Prasad (2020), we applied Gaussian noise distributions $N(0, 0.01)$ or $N(0, 0.05)$ to each feature of the benchmark dataset. Then, we tested all the well-known forecasting methods on these datasets with the noise following $N(0, 0.01)$ or $N(0, 0.05)$. Notice that, the number of uncertain points and related features in the benchmark datasets with artificial noise respectively become as many as 12000 and 97, which are much higher than those of our dataset in the JD.com case. The MAPE values of all tested methods on these benchmark datasets with noise following $N(0, 0.01)$ and $N(0, 0.05)$ are provided in Tables K2 and K3, respectively. It is clear that the proposed DHPP model dominates other tested methods under these scenarios with large-scale datasets in terms of forecasting accuracy. This advantage becomes even more pronounced as the standard deviation of the noise increases. Moreover, as shown in Table K4, the computational time of proposed DHPP model on these large-scale benchmark datasets is limited and acceptable. These numerical results further demonstrate the

robustness and efficiency of the DHPP model, highlighting its ability to handle large-scale uncertain data effectively and efficiently. Moreover, the success of the DHPP model in these benchmark datasets reinforces its potential to address forecasting challenges beyond the JD.com warehouse context (even with much larger datasets), showcasing its broader applicability and generalizability in handling real-world scenarios.

Table K1 Details of UCI Benchmark Dataset Test

Data set	Slump	Autoprice	Machine	MPG	Housing	Traffic	Cargo	Lattice
Number of points	103	159	209	392	506	2101	3942	12000
Number of dimensions	8	15	7	8	14	48	98	39

Table K2 MAPEs of All Tested Methods on Benchmark Datasets with Noise Following $N(0, 0.01)$

Dataset	**LR**	**SVR_G**	**NN**	**RF**	**LASSO**	**RR**	**XGBoost**	**MLR**	**QSVR**	**DHPP**
Slump	0.0782	0.0497	0.2318	0.1099	0.0781	0.0768	0.0919	0.0592	0.0931	**0.0126**
Autoprice	0.2032	0.1525	0.1885	0.1610	0.2008	0.1947	0.1620	0.1289	0.3077	**0.1206**
Machine	0.9676	0.4073	0.5771	0.4740	0.6502	0.7191	0.4609	0.7561	0.5849	**0.4571**
MPG	0.1262	0.0904	0.0997	0.0982	0.1238	0.1233	0.0962	0.1222	0.0998	**0.0850**
Housing	0.1872	0.1322	0.1676	0.1531	0.1861	0.1809	0.1423	0.1686	0.1371	**0.1289**
Traffic	0.4241	0.3827	0.2935	0.2115	0.9783	0.1581	0.1713	0.1097	0.0898	**0.0126**
Cargo	0.7865	0.9437	0.9549	0.6463	0.8200	0.8131	0.5999	0.6222	0.5424	**0.5045**
Lattice	0.4058	0.5003	0.5538	0.5077	0.5063	0.4052	0.4366	0.4835	0.4841	**0.2008**
Mean	0.3973	0.3323	0.3834	0.2952	0.4429	0.3339	0.2702	0.3023	0.2964	**0.1903**

Table K3 MAPEs of All Tested Methods on Benchmark Datasets with Noise Following $N(0, 0.05)$

Data set	**LR**	**SVR_G**	**NN**	**RF**	**LASSO**	**RR**	**XGBoost**	**MLR**	**QSVR**	**DHPP**
Slump	0.0793	0.0495	0.0644	0.1072	0.0785	0.0772	0.0829	0.0609	0.0935	**0.0194**
Autoprice	0.2032	0.1525	0.2025	0.1634	0.2008	0.1947	0.1631	0.1165	0.3080	**0.1114**
Machine	0.9678	0.4025	0.5851	0.4876	0.6502	0.7193	0.4741	0.8937	0.5830	**0.4584**
MPG	0.1262	0.0897	0.1085	0.0996	0.1237	0.1233	0.1022	0.1162	0.1009	**0.0883**
Housing	0.1876	0.1573	0.1673	0.1544	0.1865	0.1811	0.1449	0.1695	0.1471	**0.1376**
Traffic	8.5264	0.5113	0.3974	0.3156	1.0000	0.2572	0.3288	0.9995	1.0000	**0.2124**
Cargo	0.7815	0.9437	0.9844	0.6889	0.8200	0.8131	0.5771	0.9222	0.5844	**0.5473**
Lattice	0.4074	0.5013	0.5586	0.5120	0.5056	0.4068	0.4427	0.4801	0.5216	**0.2032**
Mean	1.4099	0.3510	0.3835	0.3161	0.4457	0.3466	0.2895	0.4543	0.4328	**0.2223**

Table K4 Average training and testing time of DHPP on benchmark datasets

Data set	Slump	Autoprice	Machine	MPG	Housing	Traffic	Cargo	Lattice
Training time (s)	0.82	13.88	0.67	7.61	27.38	650.14	3422.37	2869.42
Testing time (s)	0.011	0.014	0.012	0.014	0.022	0.035	0.041	0.063

Appendix L: Sensitivity Analysis of Selected Features

We conducted additional numerical experiments to investigate how different feature groups affect the forecasting performance of the proposed DHPP model. Specifically, based on the feature selection using the three machine learning methods, we segmented the 20 features into three groups, reflecting each feature's ranking for predicting the picking time (see Table L5) and the performance inconsistency (see Table L7). The numerical results are shown in Tables L6 and L8. Note that all 20 features collectively are referred to as Group O in these tables.

Based on Tables L6 and L8, we draw the following observations. First, in all numerical experiments, using Group I features yields more accurate forecasts than Group II, while Group II consistently outperforms Group III. This

finding highlights that features in Group I are the most critical, while those in Group III are the least significant. This conclusion aligns with the feature importance rankings derived from the three machine learning methods, as shown in Tables B4 and B6. Second, using feature Group I results in a better performance than using Group II or III when predicting the picking time for some product types or predicting the performance inconsistency in certain shifts. Third, using all 20 features (Group O) consistently outperforms using any individual feature group, emphasizing the importance of incorporating the complete feature set for accurately predicting picking performance.

Table L5 Segmentation of Features Utilized for Predicting the Picking Time

	Feature Selection
Group I	x_{19}, x_3, x_1, x_5, x_8, x_4
Group II	x_{10}, x_{15}, x_{14}, x_{11}, x_{17}, x_6, x_2
Group III	x_{20}, x_{13}, x_7, x_{16}, x_{12}, x_9, x_{18}

Table L6 MAPEs of DHPP by Using Different Feature Groups for Predicting the Picking Time

Shift	Type (Zone)	Group I	Group II	Group III	Group O
Early morning	Type1	0.1150	0.1434	0.1668	**0.0912**
	Type2	0.1322	0.1367	0.1624	**0.1292**
	Type3	0.1298	0.1354	0.1793	**0.1116**
	Type4	0.1594	0.1754	0.1844	**0.1296**
	Type5	0.0956	0.1196	0.1351	**0.0717**
	Mean	0.1264	0.1421	0.1656	**0.1067**
Daytime	Type1	0.2005	0.2018	0.2074	**0.1884**
	Type2	0.2119	0.2050	0.2119	**0.1925**
	Type3	0.1304	0.1503	0.1705	**0.1119**
	Type4	0.1269	0.1400	0.1520	**0.1143**
	Type5	0.1957	0.2030	0.2148	**0.1612**
	Mean	0.1731	0.1800	0.1913	**0.1537**
Nighttime	Type1	0.1590	0.1605	0.1689	**0.1442**
	Type2	0.1273	0.1760	0.1766	**0.0754**
	Type3	0.1652	0.1781	0.1896	**0.1412**
	Type4	0.1604	0.1925	0.1998	**0.1463**
	Type5	0.1760	0.1856	0.2027	**0.1556**
	Mean	0.1576	0.1785	0.1875	**0.1325**

Table L7 Segmentation of Features Utilized for Predicting the Performance Inconsistency

	Feature Selection
Group I	x_{19}, x_{16}, x_5, x_3, x_{12}, x_{17}
Group II	x_4, x_{14}, x_6, x_8, x_7, x_{15}, x_{13}
Group III	x_{10}, x_1, x_{20}, x_2, x_9, x_{11}, x_{18}

Table L8 MAPEs of DHPP by Using Different Feature Groups for Predicting the Performance Inconsistency

Shift	Group I	Group II	Group III	Group O
Early morning	0.0540	0.1115	0.1922	**0.0360**
Daytime	0.0576	0.1200	0.1420	**0.0395**
Nighttime	0.0655	0.0874	0.1348	**0.0374**
Mean	0.0590	0.1063	0.1563	**0.0376**

Appendix M: Calculation of the Payback Period of RFC

The calculation of the payback period of RFC aligns with JD.com's method for evaluating investments in intelligent warehousing, as detailed in Qin et al. (2022). Specifically, we define $f = pN - C_0$, where $p = p_2 - p_1$ denotes the cost difference per order between traditional (p_1) and intelligent warehouse (p_2), and N is the total number of orders completed by the intelligent warehouse over five years, and C_0 is the investment cost for the intelligent warehouse (including construction, equipment, labor, management, energy consumption, packaging, and so forth). JD.com would consider investing in intelligent warehousing if $f > 0$ and refrain if $f \leq 0$.

In the calculation process, we take $N = n_0 * 365 * 5$, where $n_0 = 24089$, which is the actual average daily order completion number for the robotic warehouses of JD.com. Due to trade confidentiality considerations, we assume $10 \leq p \leq 50, 10^8 \leq C_0 \leq 2.25 * 10^{10}$. Applying our DHPP model and the assignment model, the 7.5% increase in

employee picking volume boosts n_0 to $n_1 = (1+7.5\%)n_0$, thereby increasing the total order volume N. As illustrated in Figure M1, this results in a shift upward in the curve for our proposed model compared to the benchmark scenario (without the DHPP model and the assignment model). The area under the curve, which represents the investment feasibility ($f > 0$), indicates that a higher investment cost can now be accommodated within the five-year payback period. If we translate this into a reduction in the payback period, the period shortens to 4.65 years, calculated as $\frac{5}{1+7.5\%}$, where 7.5% represents the picking productivity improvement.

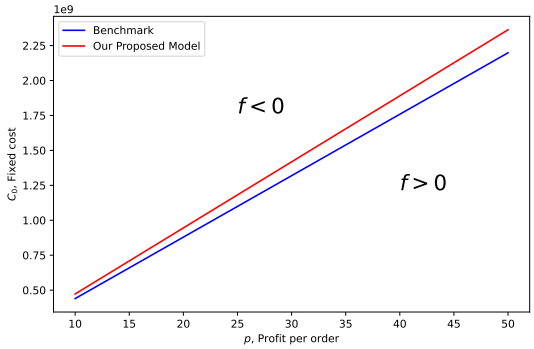

Figure M1 Decision-making Chart for Investment in Intelligent Warehousing.

Appendix N: Details of Picking Volumes of the Five Types of Products

Using data spanning from May 6th, 2019 to January 2nd, 2020, Figure N2 provides the details of picking volumes completed by all the warehouse workers across different shifts of the day: early morning (00:00 a.m.- 08:00 a.m.), daytime (08:00 a.m.- 16:00 p.m.), and nighttime (16:00 p.m.- 00:00 a.m.). The five colors in each shift represent the five product types, with the height of each colored bar indicating the picking volume at that type. The percentages on top of the bars represent the proportion of picking volume at each type with respect to the total volume during that shift. From Figure N2 we can observe that the ratio of each type of product in the three shifts is stable and that the total volume of picked products is highest in the daytime and lowest in the early morning. To statistically verify this observation and rule out the influence of this objective factor, we performed Spearman correlation analyses on the picking volumes of the five product types across the three time periods. Spearman correlation measures the monotonic relationship between datasets and is suitable when focusing on the trend shape rather than on specific numerical values. The results show that the pairwise Spearman correlation coefficients among the three groups are all 0.99 with p-values less than 0.001, indicating that the trend patterns of the three datasets are almost identical. This result indicates that, although the total number of picked products may differ, they are highly similar in terms of change direction and pattern, exhibiting a significant monotonic relationship.

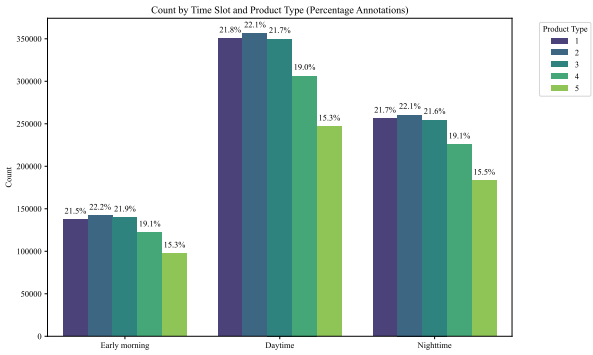

Figure N2 Picking Volume Distribution by Shift and Product Type

Appendix O: Illustration of the Randomness of the Feature Values

To illustrate the randomness of the feature values, we plot the distribution of several independent features (i.e., picking time and performance inconsistency) and two dependent features of a representative picker, fitting them with a normal distribution (see Figure O3). We then recorded the percentage of points falling within a certain range (see Table O9). As shown from the figure and table, the majority of data points lie within one standard deviation above or below the mean, and most data points fall within two standard deviations. This observation highlights the random nature of these independent and dependent variables.

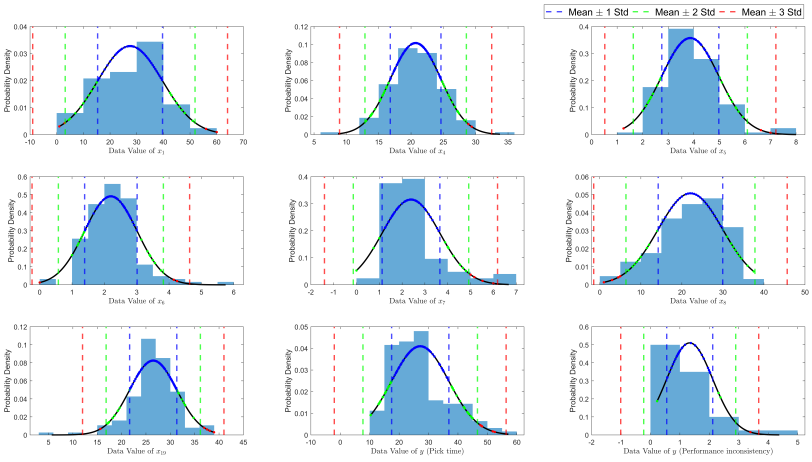

Figure O3 Data Distribution Features with Normal Distribution Fit and Standard Deviation Lines

Table O9 Percentage (%) of Data Points Within Acceptable Range with Normal Distribution Fit

Scale	x_1	x_4	x_5	x_6	x_7	x_8	x_{19}	y(Pick Time)	y(Performance inconsistency)
Mean±1 Std	68.8	70.4	73.6	78.4	81.4	68.0	78.4	68.0	87.5
Mean±2 Std	96.8	96.8	96.0	95.2	92.4	96.8	94.4	95.2	95.0
Mean±3 Std	100	98.4	98.4	98.4	95.6	100	94.8	95.6	97.5

Appendix P: The impact of Performance Inconsistency on Assignment Performance

In this section, we conduct the experiment to investigate how the performance inconsistency affect the assignment performance. Specifically, we set all positively predicted G_{it} to the minimum predicted value across all workers with positive G_{it}, i.e., $G_{it} = min\{\text{positive } G_{it}\}$, during the optimization phase, while using their actual predicted G_{it} values during simulation. The rationale behind this setting is as follows: assigning the minimum G_{it} value leads the assignment optimization model to incorrectly assume that all workers become more efficient during consecutive shifts, resulting in scheduling excessive consecutive shifts. In practice, many pickers experience reduced efficiency due to fatigue from excessive continued work. This misalignment between incorrect prediction and reality creates a significantly negative impact on system performance.

Table P10 illustrates the significant impact of G_{it} prediction accuracy on the performance of different optimization methods. We find that while the DHPP method continues to outperform QSVR and FCFS, its effectiveness is substantially weakened under inaccurate G_{it} assumptions. Specifically, the improvement of the DHPP model in picking volume drops from 7.5% to 1.7% (a relative decrease of 77.3%) and the reduction in remaining orders drops from 14.2% to 3.2% (a relative decrease of 77.5%). Note that the QSVR model suffers the most dramatic improvement decline—its picking volume shifts from a 0.5% picking volume improvement to a 4.8% decline, and remaining orders drops from a 0.9% improvement to a 9% decline. These results underscore that accurate G_{it} prediction is important to obtain an efficient assignment schedule.

Table P10 Simulation Results

	Benchmark	Optimization models		Mean percentage change (%)	
	FCFS	QSVR	DHPP	QSVR	DHPP
Average items per picker	223.5 (1.69)	213.0 (0.48)	227.4 (0.52)	-4.8 (0.63)	1.7 (0.67)
Remaining items	17783.0 (253.83)	19365.1 (72.5)	17198.1 (77.63)	9.0 (1.37)	-3.2 (1.2)

Note: Standard errors are shown in parentheses.

In conclusion, our analysis highlights the substantial benefits of accurate G_{it} prediction, revealing that prediction inaccuracies can reduce system effectiveness by over 77%.

Appendix Q: Statistical Analysis of Picker Performance Inconsistency

To statistically validate these observations, we conduct pairwise t-tests, which reveal significant differences between each shift. Specifically, we observe a significant decrease in picking time ($t = 50.004, p < 0.001$) from early morning to daytime, suggesting improved efficiency as the day progresses. A similar trend was observed from daytime to nighttime, with a significant decrease ($t = 22.577, p < 0.001$) in picking time. However, efficiency dropped significantly from Nighttime to the following Early morning slot, reflected in a substantial increase in picking time ($t = -62.612, p < 0.001$). We also performed a one-way ANOVA to check the picking efficiency across early morning, daytime, and nighttime shifts. The results also reveal a highly significant difference in the average picking time across the shifts ($F = 2161.732, p < 0.001$), indicating considerable variation in efficiency throughout the day.

Appendix R: Equivalent Mixed 0-1 Integer Programming Reformulation of Problem (11)

Problem (11) can be reformulated into the following equivalent mixed 0-1 integer programming.

$$\begin{aligned} \max_{\boldsymbol{\tau},\boldsymbol{\chi}} \quad & f(\boldsymbol{\tau},\boldsymbol{\chi}) \\ \text{s.t.} \quad & \text{constraints } (11.1)-(11.3) \\ & \chi_{ijt} \in \{0,1\}, \tau_{it} \in \{0,1\}. \end{aligned} \tag{13}$$

where,

$$\begin{aligned} f(\boldsymbol{\tau},\boldsymbol{\chi}) = \max_{\boldsymbol{a},\boldsymbol{z},\boldsymbol{r}} & \sum_{t=1}^{T}\sum_{l=1}^{L}\rho_{tl}\sum_{j=1}^{J}\sum_{p=1}^{P}\sum_{i=1}^{A}\frac{(a_{pijtl}-G_{it}r_{pijtl})}{M_{ipt}} \\ \text{s.t.} \quad & \text{constraints } (11.4)-(11.8) \\ & r_{pijtl} \le \Omega\tau_{it}, && \forall p\in[P], i\in[A], j\in[J], t\in[T], l\in[L], \\ & r_{pijtl} \le a_{pijtl}, && \forall p\in[P], i\in[A], j\in[J], t\in[T], l\in[L], \\ & r_{pijtl} \ge a_{pijtl} - \Omega(1-\tau_{it}), && \forall p\in[P], i\in[A], j\in[J], t\in[T], l\in[L], \\ & r_{pijtl} \ge 0, && \forall p\in[P], i\in[A], j\in[J], t\in[T], l\in[L], \\ & z_{pjtl}\in\{0,1\}, a_{pijtl}\ge 0. \end{aligned}$$