



A reinforcement learning model for route allocation optimization: A case study of C-130H transport aircraft in the Royal Thai Air Force

Patikorn Anchuen¹, Phummipat Daungklang², Nuntipat Phisutthangkoon², Nattawat Tanomchad² and Hatsadin Jantaboon^{1*}

¹Office of Graduate Studies, Navaminda Kasatriyadhiraj Royal Air Force Academy, Bangkok 10220, THAILAND

²Department of Computer Science, Navaminda Kasatriyadhiraj Royal Air Force Academy, Bangkok 10220, THAILAND

*Corresponding author: hatsadin_ja@rtaf.mi.th

ABSTRACT

This study aims to develop a route-allocation model for transport aircraft in the Royal Thai Air Force's (RTAF) air logistics operations. A Reinforcement Learning (RL) approach was applied to optimize resource allocation by determining the most suitable routes based on cargo capacity distribution, thereby reducing operational flight distances and, consequently, the frequency of aircraft maintenance. This research was conducted in a simulated environment using domestic air transport data as a reference for C-130H transport aircraft of Squadron 601, Wing 6, RTAF. Experimental results show that the developed model significantly improves air transport operational efficiency in support and service provision, facilitating network operations by reducing flight distances and increasing process continuity under varying cargo capacity conditions compared to current practices. Ultimately, this contributes to the improvement and sustainability of defense operations. The proposed work schedule also demonstrates adaptability to dynamic operational constraints and changing demand.

Keywords: Air logistics, Route allocation, Reinforcement learning, C-130H transport aircraft

INTRODUCTION

Tactical air transport and logistics play a vital role in military operations and national security, particularly in the Royal Thai Air Force, whose primary mission is to support personnel and equipment to enable effective operations. Under the Royal Thai Air Force's 20-Year Strategy (2018-2037) (Revised 2020), Strategic Issue 2 emphasizes strengthening national defense capabilities and readiness, including the development and improvement of the logistics system to meet international standards [1]. However, the air logistics process still faces several challenges, especially in reducing transport distances and times while increasing safety and operational efficiency. Although air transport is a fast and highly flexible option, it incurs high costs, particularly in fuel, which is a major factor affecting the overall operating costs [2]. The C-130H short-body tactical transport aircraft, based at Squadron 601, Wing 6, has a high fuel consumption rate due to its size and the volume of equipment it carries [3]. Therefore, efficient aircraft planning and resource allocation have become essential. Management is a key factor directly related to the Royal Thai Air Force's combat readiness, agility, and long-term sustainability, achieved by reducing fuel consumption and increasing operational efficiency.

At present, route and fleet management optimization have attracted considerable attention in both commercial transportation and security-related applications, as these problems are inherently complex and directly impact operational expenses and resource utilization efficiency [4, 5]. Many studies are currently being conducted on the original problem using the Vehicle Routing Problem (VRP) method, which aims to find the best way for multiple vehicles to deliver goods to different locations, taking into account time, distance, and load capacity constraints [6, 7]. The Pick-up and Delivery Problem (PDP) is a more modern version of the VRP. It can handle situations where a ministry has to pick up and drop off loads on the same route, which means moving government equipment and people and keeping track of changing needs [8, 9].

VRP and PDP in many research studies can efficiently solve problems within a limited time frame by applying mathematical methods such as Integer Linear Programming (ILP) and Mixed-Integer Programming (MIP) [10], as well as heuristics and metaheuristic methods such as Genetic Algorithm (GA), Ant Colony Optimization (ACO), and Particle Swarm Optimization (PSO) [11, 12]. However, these methods still have limitations when problems are large or exhibit complex dynamics, such as real-

time-changing demands, resource constraints, and uncertain environments.

Mathematical optimization and heuristic algorithms provide efficient routing at a fundamental level, but they often have limitations when faced with adaptive transportation situations, such as fluctuating load demands, time constraints, and constantly changing resource conditions. The resulting decisions were not made promptly and were not adaptable in actual work situations. [13] In response to these limitations, recent research has turned to more learning approaches that leverage information and environmental interactions to enhance system adaptability, particularly through the application of machine learning (ML) and deep learning (DL) techniques for decision modeling [14]. This research, therefore, applies Deep Reinforcement Learning (DRL) technology to the air logistics planning process to improve operational efficiency and increase the speed and safety of transporting personnel and equipment from one point to another, reducing distance, time, and resource allocation in a bidirectional manner for simultaneous pick-up and delivery of goods.

MATERIALS AND METHODS

Route allocation for C-130H air transportation is a complex combinatorial optimization problem constrained by fuel capacity, payload, and flight range, and can be modeled as VRP or PDP. Traditional approaches have been widely applied but often struggle in dynamic and uncertain environments [7, 12]. As a result, reinforcement learning has gained attention as a promising alternative for achieving more adaptive and effective route optimization.

1. Deep reinforcement learning for routing problems

Reinforcement Learning (RL) enables an agent to learn optimal decision-making policies by interacting with an environment and maximizing long-term cumulative rewards. In route allocation problems, the agent performs sequential route selection under operational constraints. Due to the large state and action spaces involved, Deep Reinforcement Learning (DRL), particularly the Deep Q-Network (DQN), is well-suited for this task [15, 16]. By approximating state-action Q-values with neural networks, DQN enables adaptive, data-driven route optimization under dynamic operational conditions.

2. Data collection and preparation

The goal of this study was to create a deep reinforcement learning model capable of routing the C-130H aircraft. A Deep Q-Network (DQN) was trained and tested in a simulated environment incorporating historical flight data. The results obtained under different operating conditions provide insights into the model's performance, robustness, and adaptability,

while demonstrating its potential applicability to real-world air logistics missions.

Each table or figure must include a brief description of the results being presented and other necessary information in a legend. Number tables consecutively in accordance with their appearance in the text, and place any table notes below the table body. Be sparing in the use of tables and ensure that the data presented in them does not duplicate results described elsewhere in the article. Please avoid using vertical rules and shading in table cells. Figures can be sized between column and page widths if the author chooses; however, it is recommended that figures are not sized less than column width unless necessary. When texts are typed completely, authors can perform text formatting. Style names are summarized and tabulated below.

This study's data collection is based on flight operation statistics of the C-130H transport aircraft operated by Squadron 601, Wing 6, RTAF [1], and it analyzes domestic military airlift patterns using five years of historical data [2]. Flight routes, airspeed, altitude, flight hours, payload weight, and fuel consumption are among the information gathered. According to the analysis, the most popular routes start at Wing 6, Bangkok Don Mueang, and connect to several other air wings, such as Wing 7, Surat Thani; Wing 21, Ubon Ratchathani; Wing 23, Udon Thani; Wing 41, Chiang Mai; Wing 56, Songkhla; Wing 1, Nakhon Ratchasima; and Wing 46, Phitsanulok.

These routes serve as the foundational data for constructing the route network model, as illustrated in Figure 1, with inter-wing distances depicted in Figure 2.

The payload capacity of the C-130H transport aircraft is defined by a maximum gross weight of 80,000 pounds [3]. This weight, which includes about 32,000 pounds of fuel, gives military equipment an effective payload capacity of 48,000 pounds [4]. Five primary pallets (Pallets 1–5), each with a maximum capacity of 8,500 pounds, and an auxiliary pallet (Pallet 6) with a maximum capacity of 5,500 pounds make up the useful payload. The fuel load of 32,000 pounds is illustrated in Figure 3.

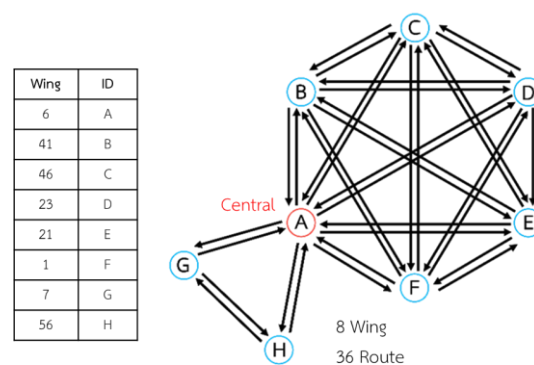


Figure 1 Domestic air wing flight route network.

Distance(Km)								
ID	A	B	C	D	E	F	G	H
A		570	320	450	480	195	555	775
B	570		260	500	780	540		
C	320	260		275	520	335		
D	450	500	275		325	280		
E	480	780	520	325		300		
F	195	540	335	280	300			
G	555							280
H	775						280	

Figure 2 Distances between air wings.

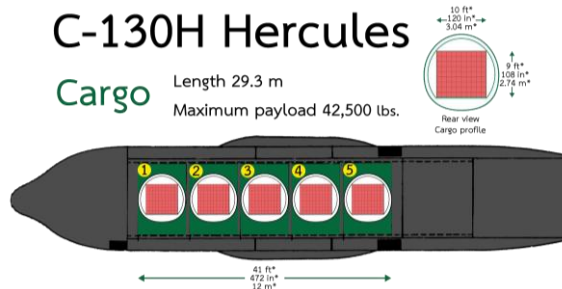


Figure 3 Number of pallets that can be carried on the C-130H aircraft.

The C-130H aircraft is capable of operating for a maximum flight duration of approximately 5 hours and 42 minutes, covering a distance of about 2,662 kilometers [6]. In addition, the aircraft can hold for approximately 1 additional hour in emergencies, in accordance with flight safety regulations [7]. The fuel consumption rate for JP-8 or Jet A-1 fuel is approximately 6,000 pounds during the first hour of operation, including engine start procedures, and approximately 4,500 pounds per hour during subsequent flight phases. An additional fuel reserve of 4,500 pounds is maintained for emergency contingencies [8].

3. Model development

Problem Formulation This study formulates the airlift route allocation problem as a sequential decision-making process under operational constraints. The objective is to minimize transportation cost while maximizing cargo delivery efficiency. The optimization objective is defined as:

$$\min J = \sum_{t=1}^T (\alpha D_t + \beta F_t - \lambda P_t) \quad (1)$$

D_t : total flight distance at time step t

F_t : fuel consumption

P_t : number of pallets delivered

α, β, λ : weighting coefficients

For reinforcement learning, the reward function is defined as:

$$r_t = \lambda P_t - \alpha D_t - \beta F_t - \eta C_t \quad (2)$$

C_t : penalty for constraint violations (e.g., payload overflow, range exceedance)

η : penalty coefficient

The system state at time step t is defined as:

$$s_t = [P_1, P_2, \dots, P_8, l_t, c_t, d_t] \quad (3)$$

P_i : number of pallets at air wing i

l_t : current aircraft location

c_t : remaining payload capacity

d_t : remaining allowable flight distance

The action $a_t \in A(s_t)$ represents a feasible round-trip airlift route sequence under operational constraints. Each action corresponds to a sequence of routes such as:

$$A \rightarrow B \rightarrow F \rightarrow C \rightarrow E \rightarrow A \quad (4)$$

The action space is dynamically constrained by payload capacity, flight range, and available pallets.

The state transition is governed by:

$$s_{t+1} = f(s_t, a_t) \quad (5)$$

where the function $f(\cdot)$ updates the system state based on cargo pick-up and delivery, aircraft movement, and resource consumption.

The reward function is defined as:

$$r_t = \beta_1 P_t - \beta_2 D_t - \beta_3 F_t - \beta_4 C_t \quad (6)$$

where C_t represents penalties for constraint violations such as payload overflow or exceeding flight range.

The model is subject to the following constraints:

- Payload capacity: $\sum P_t \leq 5$ pallets per flight
- Flight range constraint: $\sum D_t \leq D_{max}$
- Fuel constraint: $\sum F_t \leq F_{max}$

This formulation ensures that routing decisions remain feasible and aligned with real-world airlift operations.

Deep reinforcement learning is used to solve routing problems in air transport operations for C-130H aircraft, accounting for constraints such as flight distance, payload capacity, and fuel consumption [9]. This approach allows the model to learn how to make decisions by interacting with the working environment [10]. Deep Q-Network (DQN), which uses a deep neural network to estimate the Q-value function, enables the model to learn complex decision policies through iterative interactions with the environment [11]. Previous studies have demonstrated the effectiveness of this algorithm in solving routing problems and improving logistics [12, 13].

The data collected in the preceding section are transformed into a flight route network representation of air wings, consisting of a total of 8 nodes representing air wings responsible for pick-up and delivery operations within a single mission round. This configuration results in 36 possible flight routes, as illustrated in Figure 1 [14]. Wing 6 is designated as the central node, while the remaining air wings are treated as transit nodes to reflect the actual operational patterns of the squadron [17]. Route characteristics, including

flight distance, flight duration, and fuel consumption between air wings, are used as input variables to construct the simulated dataset for model training [15]. These variables play a critical role in evaluating the effectiveness of decision-making for each flight route [18].

Prior to model training, an initial data filtering

process based on reward scores is performed to select actions that are most appropriate and aligned with the research objectives. This process helps reduce problem complexity and improve the stability of the learning process [19]. As a result, the model is better able to identify efficient routes, as summarized in Table 1.

Table 1 C-130H flight operation data.

Route	Flight time (Minute)	Fuel Consumption (lbs.)	Distance (Km.)
BA	73	6975	570
CA	46	4600	320
CB	39	3900	260
DA	64	6300	450
DB	84	7800	500
DC	44	4400	275
EA	66	6450	480
EB	98	8850	780
EC	70	6750	520
ED	47	4700	325
FA	33	3300	195
FB	84	7800	540
FC	51	5100	335
FD	42	4200	280
FE	44	4400	300
GA	72	6900	555
HA	96	8700	775
HG	41	4100	280

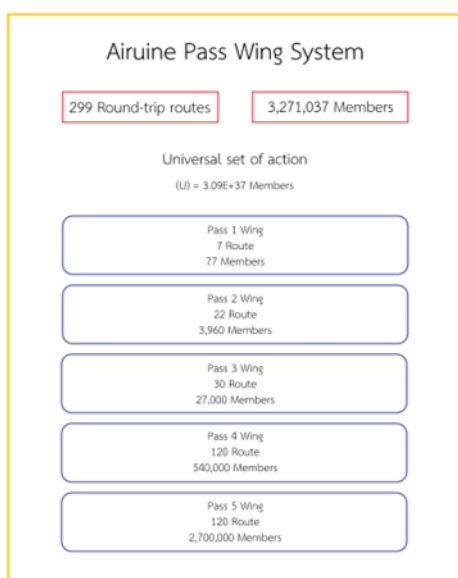


Figure 4 Route Classification Structure for DRL.

Prior to model training a data filtering process is conducted to select actions that correspond to flight routes yielding the highest rewards. The selection criteria are based on reward scores, as shown in Figure 4. A neural network architecture is used to estimate the Q-value function, which determines whether the action is optimal for each system state [21]. The loss function, which measures the difference between

predicted and actual rewards, is what the network's weights are always changing during training to minimize. This process enables the model to learn more accurate and effective decision-making policies under varying operational conditions.

After training with the DQN model, the Q value for each route can be automatically calculated, and the route yielding the highest overall reward score can be selected, taking into account factors such as payload capacity, flight distance, and fuel consumption. The results are then compared with those obtained from traditional route-allocation methods during the testing phase to verify the performance of the proposed model.

The proposed model employs a Deep Q-Network (DQN) to approximate the action-value function $Q(s,a)$. The neural network consists of two fully connected hidden layers, each containing 200 neurons. Rectified Linear Unit (ReLU) activation functions are applied to the hidden layers, while a linear activation function is used in the output layer.

The model is trained using experience replay with a memory buffer size of 10,000 transitions. A minibatch size of 36 is sampled uniformly from the replay memory at each training step. The discount factor is set to $\gamma=0.95$, and the learning process

follows an ϵ -greedy policy, where ϵ decreases linearly from 1.0 to 0.1 over the training period.

The network parameters are optimized using the Adam optimizer with a learning rate of 0.001. The model is trained for 100,000 iterations, and convergence is determined based on the stabilization of the Temporal Difference (TD) error.

A target network is employed to improve training stability, with parameters updated periodically every 1,000 iterations.

All experiments are conducted using the same dataset and configuration to ensure reproducibility.

In addition to the analysis of Temporal Difference (TD) error and reward trends, the performance of the proposed method was further evaluated using practical operational indicators. These include total flight distance, cargo throughput, and overall routing efficiency, which provide a clearer understanding of how the model performs in real-world air logistics operations.

The results show that the proposed method can significantly reduce the total flight distance compared to the conventional routing approach. In this study, the total distance decreases from 3,130 km to 2,445 km, representing a reduction of approximately

21.9%. This improvement suggests that the learned routing policy is able to eliminate unnecessary travel segments and produce a more efficient flight pattern.

Regarding cargo handling performance, the proposed method delivers 19 pallets in a single mission cycle. In contrast, the conventional approach is estimated to handle approximately 16 pallets due to repeated return trips and less efficient route planning. This corresponds to an increase of about 18% in cargo throughput, indicating better utilization of aircraft capacity.

Although fuel consumption is not directly modeled as an objective, reducing travel distance naturally leads to lower fuel use. This also implies reduced operating costs and aircraft usage, which are important factors in military air transport operations.

Another key observation is that the proposed method avoids repeated returns to the origin air wing, resulting in a more compact route structure. This allows multiple pick-up and delivery operations to be completed within a single mission, improving both time efficiency and resource utilization.

A summary of the comparison between the conventional method and the proposed approach is presented in Table 2.

Table 2 Performance comparison.

Method	Distance (km)	Reward	Pallets Delivered	Improvement
Conventional	3,130	27,870	16	baseline
Proposed (DQN)	2,445	16,555	19	-21.9% Distance and +18.8% Pallets Delivered

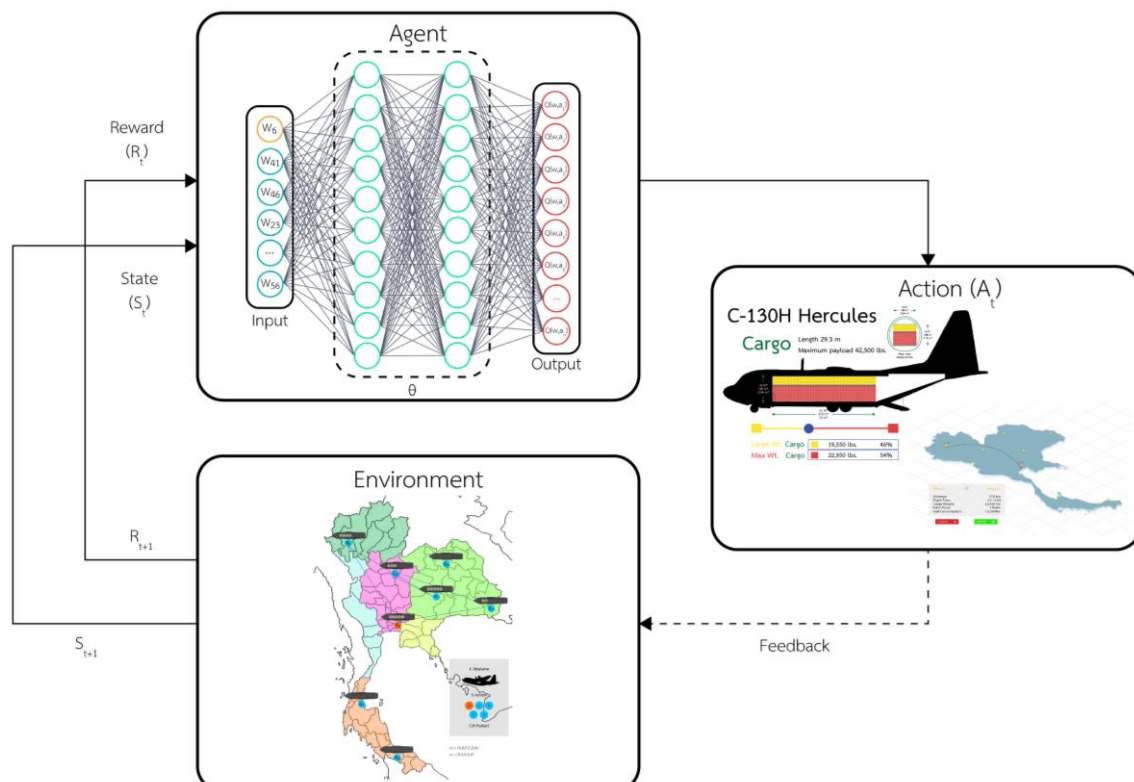


Figure 5 DQN architecture.

4. Model testing and performance evaluation

To assess the efficacy of deep reinforcement learning models, experiments were conducted in a simulated environment using operational data from either the 601st Wing or the 6th Wing of the Royal Thai Air Force. There were 36 air transport routes and 40 pallets of military gear. These data were employed as the testing dataset for both model training and performance evaluation of the decision-making process [22, 23]. During training, key neural network parameters within the DQN algorithm—including the number of training episodes, learning rate, and network depth—were configured to enable comparative analysis across three scenarios: Case I (learning rate = 0.05), Case II (learning rate of 0.07), and Case III (learning rate of 0.10) [24, 25].

The experimental results demonstrate that the model effectively learns to reduce the discrepancy between predicted and target Q-values. As the number of training episodes increases, the Temporal Difference (TD) error shows a steady decline [26], indicating improved learning stability. The DQN architecture employed in this study is illustrated in Figure 5, and the learning procedure follows the algorithm described below:

Algorithm: Deep Reinforcement Learning

1. Initialize replay memory D
2. Initialize Q-network with random weights θ
3. For each episode = 1 to N_{ep} do
 Initialize starting state s_1
 Preprocess input to get $\phi_1 = \Phi(s_1)$
 For each time step $t = 1$ to T do
 Select action a_t using epsilon-greedy policy based on $Q(\phi_t, a; \theta)$

```

Execute action  $a_t$  and observe reward  $r_t$  and
next state  $s_{t+1}$ 
Preprocess next state to get  $\phi_{t+1} = \Phi(s_{t+1})$ 
Store transition  $(\phi_t, a_t, r_t, \phi_{t+1})$  in replay
memory D
Sample random minibatch of transitions  $(\phi_j, a_j, r_j, \phi_{j+1})$  from D
If  $\phi_{j+1}$  is a terminal state then
   $y_j = r_j$ 
Else
   $y_j = r_j + \gamma * \max_{a'} Q(\phi_{j+1}, a'; \theta)$ 
End if
Perform gradient descent step on  $(y_j - Q(\phi_j, a_j; \theta))^2$ 
End for
End for
  
```

RESULTS AND DISCUSSION

The decision policy is constructed and its implementation is discussed. The DQN algorithm constructs a decision function that tells the system what to do based on its current state. This design aims to improve route allocation for C-130H operations, accounting for factors such as time and distance and fuel consumption. The proposed framework effectively simplifies the decision space while maintaining operational relevance and efficiency [27].

To make a synthetic dataset, you need to set important variables like the route ID, the action taken, the distance traveled, and the number of pallets moved. These variables are the inputs that the DQN model [20] uses to learn. The compiled data is organized and stored in a tabular format based on route sequences, as demonstrated in Table 3.

Table 3 Round-trip distances of routes under the selected actions.

Route	Action	Distance (Km.)
1	a_1-a_{35}	1140
2	$a_{36}-a_{70}$	640
3	$a_{71}-a_{105}$	900
...	...	...
299	$a_{3248706}-a_{3271205}$	1925

Table 4 Training configuration.

Parameters	Action
Data samples	10000
The number of links	36
Total Neurons	200
Total Output Layer	300
Threshold Error	0.0001
Total Iterations	100000
Learning rate	0.05/0.07/0.10

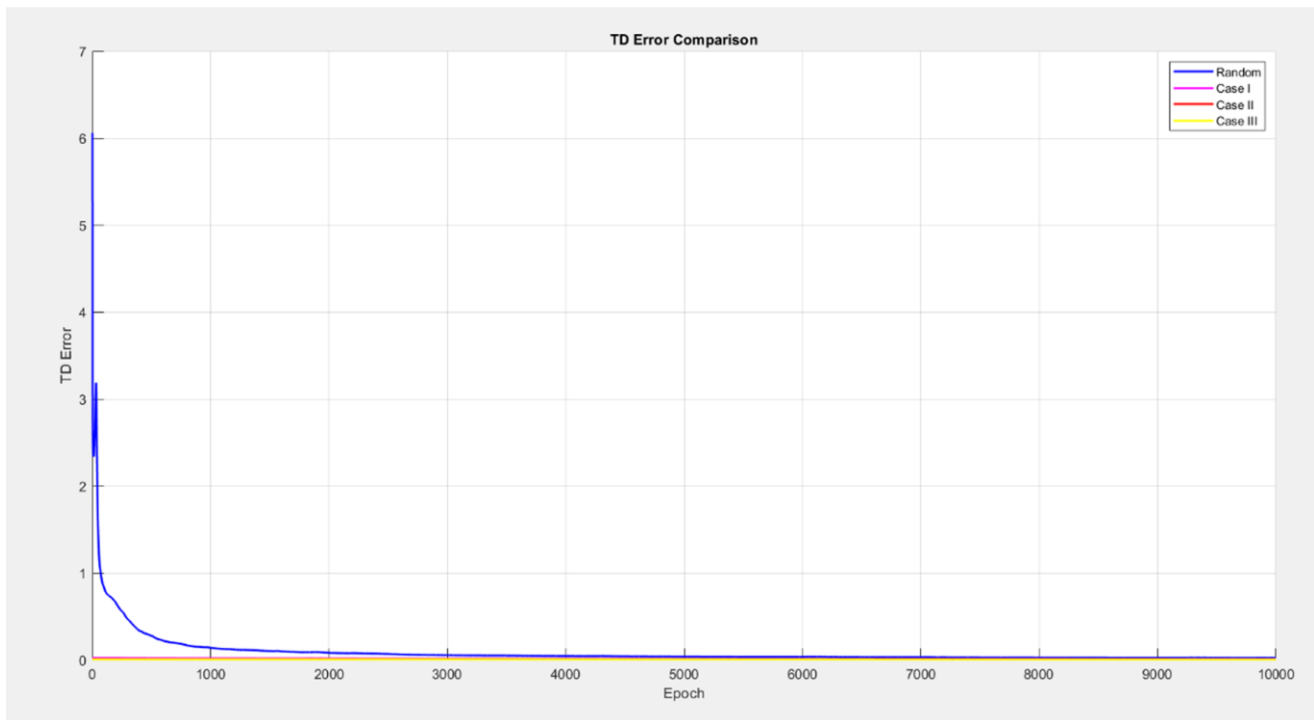


Figure 6 TD error between target values and q-values.

1. Policy construction

This stage constitutes a critical component of the deep reinforcement learning (DRL) process, in which the DQN algorithm is employed to derive an optimal decision function for allocating airlift routes for Squadron 601, Wing 6, RTAF. The primary objective is to minimize total flight distance while improving the efficiency of aviation resource utilization. The construction of a DRL framework aims to enhance various management efficiencies by enabling the algorithm to learn through interactions between an agent and the environment. This learning process is often formulated as a state-action pair, where each action receives a reward reflecting the quality of the decision at a given time step, though other formulations exist as well. The Q-value, $Q(s,a)$, representing the expected cumulative return obtained by taking action a in state s , is approximated using a neural network in accordance with the Deep Q-Network (DQN) paradigm [16, 19].

In this study, the state space consists of information on the number of pallets held at each air wing at time t , along with route distances and flight constraints. The action is to the set of available airlift routes (R1-R36) that the C-130H transport aircraft can select for operation. The DQN model incorporates an experience replay mechanism combined with the Adam optimizer to update network weights and to reduce the Temporal Difference (TD) error between predicted and target Q-values, thereby improving learning accuracy and stability [23].

The airlift activities of Squadron 601, Wing 6, comprising eight air wings, forty pallets of military equipment, and thirty-six potential routes between

origin and destination air wings, are the subject of the case study. The operational requirements of each air squadron will determine the reward for each flight route. The C-130H aircraft can accommodate up to 5 pallets per flight [28].

Given the complexity and uncertainty inherent in route allocation, this study adopts a trial-and-error learning process to identify optimal decision policies. For each experimental scenario, actions are randomly selected from the available route set and evaluated under three cases Case I (learning rate = 0.05), Case II (0.07), and Case III (0.10) to compare the learning speed and stability of the proposed model. A fully random action selection strategy is used as a baseline for performance comparison [22].

The learning rate is a key factor in how quickly the DQN neural network's weights are updated [22, 24]. To further improve the model's performance during retraining, parameters such as the number of hidden nodes, the number of training episodes, and the error threshold are set by parameters [25]. Figure 6 shows that the random baseline has a Temporal Difference (TD) error that stays pretty stable across training episodes. But this stability is due to the TD error converging to a value, not the policy converging to a value, because the random policy doesn't perform structured learning or policy improvement. So, even though the TD error in the random case looks low, the decisions and total rewards are still much lower than those from the DQN-based methods. The comparative results indicate that the random case, a conventional uninformed action selection strategy, exhibits slower learning while maintaining greater numerical stability due to the absence of policy updates. However, it does

not reach policy convergence or reward optimization [26]. On the other hand, Cases I–III show faster convergence in learning because they use different learning rates (0.05, 0.07, and 0.10, respectively). However, transient instability is observed during the initial training phase, evidenced by temporarily elevated TD errors resulting from frequent weight updates [27].

The final results reveal that as the number of training episodes increases, the algorithm consistently reduces the TD error and enhances the

effectiveness of the decision function in terms of cumulative Q-value rewards. As shown in Figure 7, the random baseline yields the lowest average reward across all states, confirming that the numerical stability observed in the TD error does not translate into effective route allocation or improved operational performance. Cases I–III ultimately achieve 100% convergence at the end of training. In contrast, the random baseline yields the lowest Q-values, as it does not account for structured route-based decision-making within the action space.

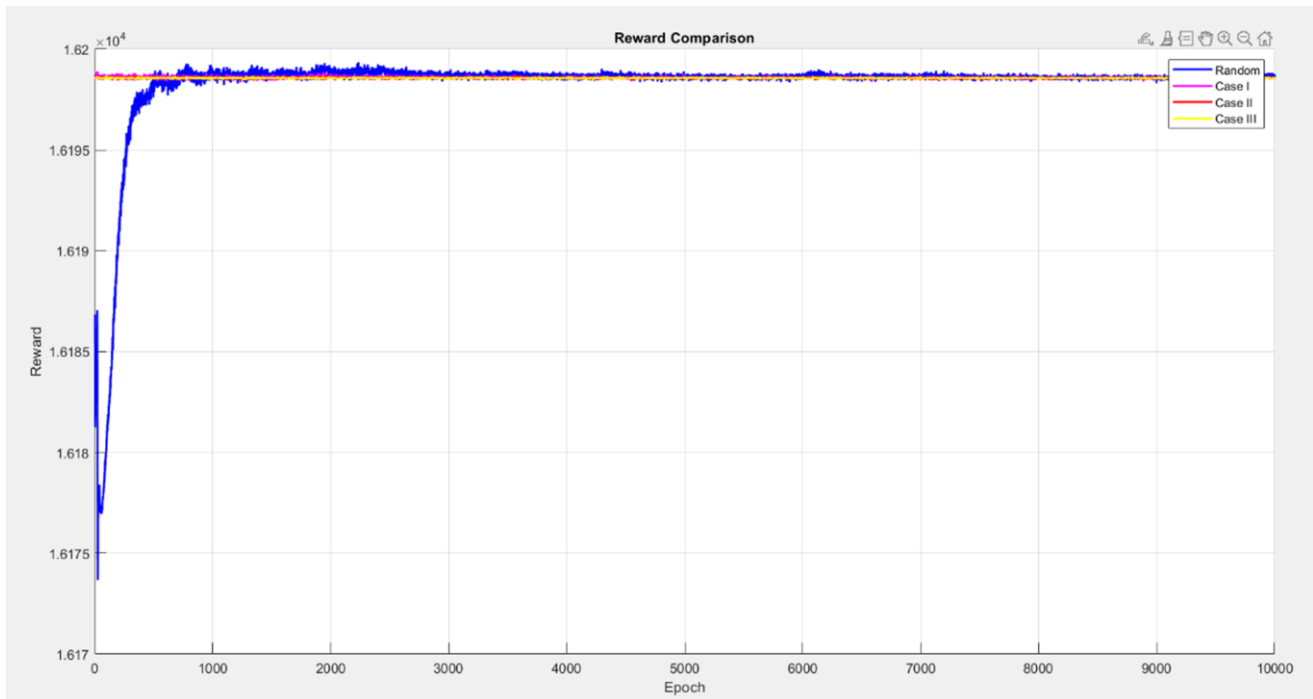


Figure 7 Average Reward for Each State.

2. Implementation

The decision function developed in the previous stage is applied to determine operational actions from the simulated system states for pallet balancing in the case study of Squadron 601, Wing 6, RTAF, which is responsible for managing C-130H airlift route allocation, as illustrated in Figure 8. The number of available pallets at each air wing is provided as input to the shared pallet management system, enabling the model to evaluate feasible transportation alternatives by computing Q-values and selecting the action that maximizes cumulative rewards under system constraints [22]. For example, the selected actions $R8 = 5$, $R29 = 4$, $R17 = 5$, $R25 = 3$, and $R4 = 2$ can be interpreted as an airlift route sequence $A \rightarrow B \rightarrow F \rightarrow C \rightarrow E \rightarrow A$. Within a single operational round, the C-130H aircraft performs pick-up and delivery operations totaling 19 pallets, covering five air wings with a total flight distance of 2,445 kilometers. The simulation results show that the proposed routing strategy significantly reduces the total flight distance compared to traditional fixed-route assignment

methods. This improves overall air transport efficiency. Moreover, the model's decision function explicitly accounts for operational constraints consistent with real-world deployment, including maximum aircraft payload capacity, flight range limitations, and mission profiles within a single airlift cycle. As a result, the proposed approach significantly enhances transportation completeness while reducing total distance-related expenses, as shown in Figure 9.

3. Discussion

From the simulation and real-world results presented in the previous section, a Deep Reinforcement Learning (DRL) model developed using the Deep Q-Network (DQN) algorithm demonstrates a significant improvement in pallet handling within the C-130H air transport system. The model learns effective decision policies under resource and time constraints by estimating the Q-value function to select the action that maximizes the cumulative reward for each system state [16]. Although the DQN algorithm shows strong potential for handling large state and action spaces, increasing the action space size

inevitably increases computational complexity. This, in turn, may result in longer training times and increased computational resource consumption, particularly in airlift routing problems characterized by numerous operational constraints [26].

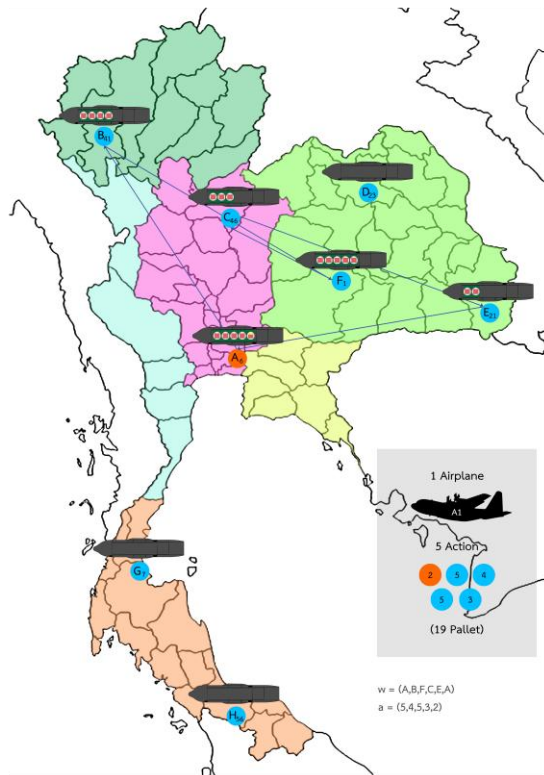


Figure 8 Application of the Decision Function for Military Airlift Operations Using C-130H Aircraft

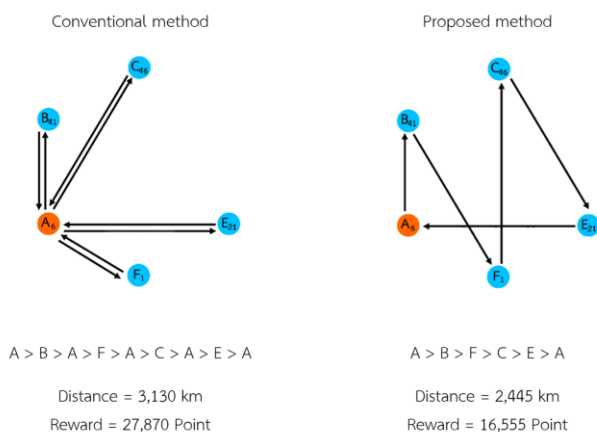


Figure 9 Comparison of Travel Distance and Reward Between the Traditional Method and the Proposed Approach for Military Airlift Operations Using C-130H Aircraft of Squadron 601, Wing 6, RTAF.

This study indicates that as the number of training episodes increases, the model's Q-values increase continuously, while the time difference error (TD) gradually decreases. This behavior reflects the stability and suitability of the DQN model's

internal parameter tuning [26]. These results are consistent with deep reinforcement learning, which aims to reduce the discrepancy between the estimated Q-value and the target Q-value once sufficient training has been completed [24]. Furthermore, applying the trained decision function to the Squadron 601, Wing 6, RTAF simulation environment has demonstrated that this model can efficiently select air transport routes that yield the highest Q-value, while maintaining operational constraints. This capability reduces operating expenses (OPEX) for military air transport missions compared to traditional routing methods.

However, an overly large action space, while potentially leading to higher Q-values, imposes a significant computational burden and slows down the learning process. Conversely, an overly limited action space can limit the model's ability to explore and identify optimal policies fully.

Therefore, designing an appropriate action-filtering strategy is crucial for balancing learning efficiency and computational complexity. This observation is consistent with accepted concepts in the field of function estimation and policy optimization within reinforcement learning frameworks [16]. Overall, the research confirms that the DQN algorithm, adapted to the operational characteristics of military air transport missions to reduce overall mission distance and increase pallet coverage under real-world operational constraints, shows promising potential for future application in military transport routing systems.

CONCLUSIONS

This study developed a pallet allocation and routing system for transportation between originating and destination air bases to address imbalances between bases. These imbalances arise from differing cargo demands, leading to overloaded aircraft or empty flights and negatively impacting the overall efficiency of the air logistics system. The proposed framework integrates a routing algorithm to appropriately filter and reduce the action space, in conjunction with a Deep Reinforcement Learning (DRL) approach—specifically the Deep Q-Network (DQN) to construct a decision function for pallet management and transportation operations. Experimental results using real operational data from Squadron 601, Wing 6, RTAF demonstrate that an optimally designed action space enables the model to learn and make decisions more efficiently. The proposed approach reduces overall flight distance, increases cargo volume, and lowers operating expenses (OPEX) compared to traditional methods. The key quantifiable performance improvements of the proposed method are summarized as follows: The total flight distance is reduced by approximately 21.9% compared to the conventional routing approach. Cargo throughput increases from an

estimated 16 pallets per mission to 19, representing an improvement of approximately 18.8%. The proposed method reduces redundant return trips to the origin air wing, resulting in a more compact, efficient routing structure. Furthermore, the learned decision-making functions can be applied to real-world logistics planning systems to support operators in making transportation decisions. Moreover, the proposed concept can be extended to other complex transportation management systems, such as land or sea logistics, which require real-time decision-making under resource constraints. Such an extension has the potential to increase operational resilience, reduce fuel consumption, and improve future transportation infrastructure. The proposed model relies on optimal factor analysis, algorithm selection, and the quality of training data. Therefore, practical implementation requires the systematic collection and management of C-130H operational data, including flight paths, speed, altitude, flight hours, and payload data. In addition, an automated, accurate data retrieval process is crucial for reducing operator workload and ensuring data continuity, enabling the model to adapt to real-world operating conditions effectively over time.

ACKNOWLEDGEMENT

This research was supported as part of the Master of Engineering program in Defense Technology, Office of Graduate Studies, Navaminda Kasatriyadhiraj Royal Air Force Academy (NKRAFA). The authors gratefully acknowledge Flt. Lt. Asst. Prof. Dr. Patikorn Anchuen for his valuable guidance, constructive comments, and continuous support throughout the research and thesis completion.

REFERENCES

1. Royal Thai Air Force. 20-Year Air Force Strategy (2018-2037). Revised 2020. Bangkok (Thailand): Royal Thai Air Force; 2020.
2. Royal Thai Air Force. Air Force Operational Plan 2023–2027. Bangkok (Thailand): Royal Thai Air Force; 2022.
3. Mouton CA, Powers JD, Romano DM, Guo C, Bednarz S, O'Connell C. Fuel reduction for the mobility air forces [Internet]. Santa Monica (CA): RAND Corporation; 2015 [cited 2026 Jan 26]. Available from: https://www.rand.org/content/dam/rand/pubs/research_reports/RR700/RR757/RAND_RR757.pdf.
4. Jayapal C, Devadharshini R, Praniis R, Praniis R. Enhancing fleet operations with a vehicle route management solution. In: Proceedings of the 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA); 2023; Coimbatore, India. p. 1-6.
5. Coppo M, Agostini M, De Matteis G, Bertolini M. Optimal management of commercial electric vehicle fleets with recharging stations and time-varying electricity prices. *Energies*. 2025;18(3): 453.
6. Bao Y. A summary and discussion on the current state of CVRP research. *Appl Comput Eng*. 2024; 115(1):35-42.
7. Reddy TS, Dhanush D, Krithin T, Sayan J. Supply chain logistics with hybrid optimization using ADMM and vehicle routing problem. In: Proceedings of the International Conference on; 2024.
8. Santos MJ, Jorge D, Bonomi V, Ramos TRP, Barbosa-Póvoa AP. Enhancing logistics through a vehicle routing problem with deliveries, pick-ups, and backhauls. *Int Trans Oper Res*. 2024.
9. Desticioğlu Taşdemir B, Özyörük B. Mathematical model for multi-depot simultaneously pick-up and delivery vehicle routing problem with stochastic pick-up demand. *Gazi Univ J Sci*. 2025;38(1): 219-35.
10. Jantaboon H, Anchuen P, Bua-ngam P. An optimization model for Air Force logistics: A case study of C-130H transport aircraft. In: Proceedings of the NCDT2025 Conference; 2025. p. 146-50.
11. Uthansakul P, Anchuen P, Uthansakul M, Khan AA. QoE-aware self-tuning of service priority factor for resource allocation optimization in LTE networks. *IEEE Trans Veh Technol*. 2020;69(1):887-900.
12. Sharma A, Srivastava S, Rautela PS, Joshi B. Metaheuristics-based routing optimization in on-chip network. In: Proceedings of the ACM International Conference; 2023.
13. Anchuen P, Uthansakul P, Uthansakul M, Poochaya S. Fleet optimization of smart electric motorcycle system using deep reinforcement learning. *Comput Mater Contin*. 2022;71(1):1925-43.
14. Poochaya S, Uthansakul P, Uthansakul M, Anchuen P, Thammakul K, Khan AA, et al. A multi-mode public transportation system using vehicular to network architecture. *Comput Mater Contin*. 2022;73(3).
15. Daungklang P, Mahavongvun A. Create a process for manually editing rapid 3D mapping datasets [Internet]. 2011 [cited 2026 Jan 26]. Available from: <https://www.diva-portal.org/smash/get/diva2:426325/FULLTEXT01.pdf>.

16. Sutton RS, Barto AG. Reinforcement learning: An introduction. 2nd ed. Cambridge (MA): MIT Press; 2018.
17. Silver D, Lever G, Heess N, Degris T, Wierstra D, Riedmiller M. Deterministic policy gradient algorithms. In: Proceedings of the International Conference on Machine Learning (ICML); 2014. p. 387-95.
18. Arulkumaran K, Deisenroth M, Brundage M, Bharath A. Deep reinforcement learning: A brief survey. IEEE Signal Process Mag. 2017;34(6):26-38.
19. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Marc G, et al. Human-level control through deep reinforcement learning. Nature. 2015;518(7540):529-33.
20. Busoniu L, Babuska R, De Schutter B, Ernst D. Reinforcement learning and dynamic programming using function approximators. Boca Raton (FL): CRC Press; 2017.
21. Watkins C, Dayan P. Q-learning. Mach Learn. 1992;8(3-4):279-92.
22. Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, et al. Playing Atari with deep reinforcement learning. arXiv preprint. 2013;arXiv:1312.5602.
23. Lillicrap TP, Hunt JJ, Pritzel A, Heess N, Erez T, Tassa Y, et al. Continuous control with deep reinforcement learning. arXiv preprint. 2015; arXiv:1509.02971.
24. Hessel M, Modayil J, Van Hasselt H, Schaul T, Ostrovski G, Dabney W, et al. Rainbow: Combining improvements in deep reinforcement learning. In: Proceedings of the AAAI Conference on Artificial Intelligence; 2018. p. 3215-22.
25. Wang Z, Schaul T, Hessel M, Hasselt H, Lanctot M, Freitas N. Dueling network architectures for deep reinforcement learning. In: Proceedings of the 33rd International Conference on Machine Learning (ICML); 2016. p. 1995-2003.
26. Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double Q-learning. In: Proceedings of the 30th AAAI Conference on Artificial Intelligence; 2016. p. 2094-2100.
27. Baneshi F, Hernandez ML. Network-level aircraft trajectory planning via multi-agent deep reinforcement learning. Comput Ind Eng. 2025;175:109057.
28. Koh S, Zhou B, Fang H. Real-time deep reinforcement learning-based vehicle routing and navigation. Appl Soft Comput. 2020;96:106694.