A Supply Chain Inventory Management Method for Civil Aircraft Manufacturing Based on Multi-Agent Reinforcement Learning

Mingjie Piao 1,*, Dongdong Zhang 1,†, Hu Lu 2 and Rupeng Li 2

1 College of Electronics and Information Engineering, Tongji University, Shanghai 201804, China
2 COMAC Shanghai Aircraft Manufacturing Co., Ltd., Shanghai 201324, China
* Correspondence: ddzhang@tongji.edu.cn
† These authors contributed equally to this work.

Abstract: Effective supply chain inventory management is crucial for large-scale manufacturing industries such as civil aircraft and automobile manufacturing to ensure efficient manufacturing. Generally, the main manufacturer makes the annual inventory management plan, and contacts with suppliers when some material is approaching critical inventory level according to the actual production schedule, which increases the difficulty of inventory management. In recent years, many researchers have focused on using reinforcement learning method to study inventory management problems. Current approaches were mainly designed for the supply chain with single-node multi-material or multi-node single-material mode, which are not suitable to the civil aircraft manufacturing supply chain with multi-node multi-material mode. To deal with this problem, we formulated the problem as a partially observable Markov decision process (POMDP) model and proposed a multi-agent reinforcement learning method for supply chain inventory management, in which the dual-policy and information transmission mechanism was designed to help the supply chain participant improve the global information utilization efficiency of the supply chain and the coordination efficiency with other participants. The experiment results show that our method has about 45% performance improvement on efficiency compared with current reinforcement learning-based methods.

Keywords: multi-agent reinforcement learning; supply chain; inventory management

1. Introduction

Due to the global distribution of factories and suppliers, the dominant supply chain model for large-scale manufacturing industries, such as civil aircraft and automobile manufacturing, is the main manufacturer-supplier model. In this model, suppliers produce parts, and the manufacturer coordinates and manages the smooth operation of the supply chain. At the same time, the main manufacturer is also responsible for the final assembly of the product. Efficient supply chain inventory management is crucial for these industries to ensure the timely material supply and production while reducing supply chain inventory costs. Many researchers have attempted to use advanced technologies to solve the problems related to this kind of supply chain inventory management.

In recent years, with the development of machine learning-related technologies, many machine learning-based methods have been applied in supply chain demand forecasting [1,2], risk management [3], transportation management [4], spare stock [5], inventory replacement [6,7], and supplier selection [8]. In traditional supply chain inventory management, it is difficult to obtain, predict or estimate the information in the supply chain accurately using traditional decision rules because it is based mainly on the experience and judgment of the inventory management personnel. Reinforcement learning-based methods are well suited for sequential decision-making problems and can learn nonlinear features in problem environments. On the one hand, it provides a new solution for
supply chain inventory management problems, such as the application to improve the
performance in perishable inventory management [9]. On the other hand, compared to
Federated Learning [10] and traditional DNN, RL method does not require a large amount
of labeled data. It only requires an initial setup of the training environment, allowing the
agent to interact with the environment and gather a large amount of data for training and
policy updates. Therefore, we chose RL as our research method. The main related works
about inventory management are listed in the Table 1. Based on the type of RL algorithm
used, research on RL-based inventory management can be classified into value-based
RL methods, policy-based RL methods, and Actor-Critic RL methods. We will take the
literature review from these three aspects.

Table 1. The main related works about inventory management with RL.

<table>
<thead>
<tr>
<th>Contributors</th>
<th>Contributed Content</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cuartas, C. [12]</td>
<td>Discussed reward function definition based on Q-Learning</td>
</tr>
<tr>
<td>Nurkasahanah, I. [13]</td>
<td>Setted minimizing costs as the optimize target</td>
</tr>
<tr>
<td>Jiang, C. [14]</td>
<td>Setted maintaining a specific target inventory level</td>
</tr>
<tr>
<td>Oroojlooy, A. [15]</td>
<td>Combined DQN and heuristic algorithms in the beer game</td>
</tr>
<tr>
<td>Kemmer, L. [16]</td>
<td>Used SARSA in the multi-node single-item mode supply chain inventory management problem</td>
</tr>
<tr>
<td>Hutse, V. [17]</td>
<td>Added the lead time in delivery</td>
</tr>
<tr>
<td>Alves, J.C. [18]</td>
<td>Constructed a four-level single material supply chain environment</td>
</tr>
<tr>
<td>Alves, J.C. [19]</td>
<td>Compared performance of some RL algorithms</td>
</tr>
<tr>
<td>Wang, H. [20]</td>
<td>Researched aviation inventory management of multiple materials under a single node</td>
</tr>
</tbody>
</table>

1.1. Inventory Management Methods with Value-Based RL

Value-based RL methods update the value function during the interaction with
the environment and make decisions based on the value function. Giannoccaro and
Pontrandolfo [11] were the first to apply RL to optimize supply chain and inventory man-
agement in 2002. In recent years, the RL methods are applied to address supply chain inven-
tory management problems due to its advantages in handling sequence decision-making
problems such as inventory management. RL-based supply chain inventory management
methods mainly involve agents learning how to effectively manage supply chain inventory
through feedback from the rewards obtained by interacting with the environment. The ob-
jective is to optimize policies to maximize profits [11], minimize costs [13,21], and maintain
a specific target inventory level [14]. Cuartas et al. [12] discussed several reward function
definition patterns based on Q-Learning algorithm for RL-based inventory management
method. A significant portion of research only relies on simple RL methods [22], such as
Q-learning [23], a classic value-based RL method. However, these simple methods face the
significant challenge of the curse of dimensionality when dealing with high-dimensional
real-world state spaces.

The Deep Q-Network (DQN) algorithm is another classic value-based RL method.
This approach is a natural extension of the Q-learning method in the field of deep learning.
Oroojlooy [15] attempted to use Deep Q-Network (DQN ) DQN to solve the beer game
problem, a fourth-level linear supply chain environment considering both deterministic
and stochastic demands. In the experiment, only one supply chain participant made deci-
sions using DQN, while the other participants followed specific heuristic algorithms. The
environment state comprised inventory levels, demands, incoming orders, and incoming
products over the past m (hyperparameter) periods. The experiment showed that from the
perspective of a single node in the supply chain, using RL methods to make decisions on
supply chain inventory management was more efficient than using heuristic methods.
1.2. Inventory Management Methods with Policy-Based RL

Policy-based RL methods update the policy function during the interaction with the environment via gradient descent and make decisions based on the policy function. Kemmer et al. [16] employed Approximate State-Action-Reward-State-Action (SARSA, a value-based RL algorithm) and three slightly different REINFORCE (a policy-based RL algorithm) algorithms on a two-level supply chain environment, extending RL methods to a multi-node single-item mode supply chain inventory management problem. The environment consisted of a factory and one to three warehouses with a horizon of 24 periods. The environment state was composed of inventory level and demand over the last two periods, and the agent’s action was factory production and product transportation. The environment reward was profit, considering warehouse maintenance costs and inventory backlogs. Based on this work Hutse [17] added lead time delivery cycles to the scenario, using DQN to handle discrete action. The environment state consisted of inventory, production, transportation, and the last \( m \) (hyperparameter) demands. The agent’s action was factory production and product transportation. The environment reward was profit considering operational costs.

Alves and Mateus [18] constructed a four-level supply chain environment for a single material in a multi-node system. They used a single-agent algorithm to solve the multi-agent problem. However, in practical problem environments involving multi-agent participation, agents often cannot access all the environment information, making such methods somewhat impractical. The lack of necessary information exchange among agents leads to unstable environments.

1.3. Inventory Management Methods with Actor-Critic RL

The Actor-Critic RL combines the advantages of both value-based and policy-based methods. It evaluates the policy function based on the value function and makes decisions relying on the policy function. Currently, most mainstream RL methods are based on this architecture. Alves and Silva [19] used a shared policy in a supply chain collaboration environment to compare the performance of different single-agent RL algorithms, such as Deep Deterministic Policy Gradient (DDPG) [24], Soft Actor-Critic (SAC) [25], and Proximal Policy Optimization (PPO) [26], and the implementation results showed that the PPO algorithm performed the best. In this study, all homogenous agents used the same policy for inventory management. However, in actual supply chains, different participants often face significantly different environmental states and cannot be simply regarded as homogeneous agents.

Barat et al. [27] attempted to use RL methods to solve inventory management problems for multiple materials. However, their approach only applies to inventory management problems for a single node. It is unsuitable for the problem of multiple nodes and materials in the scale manufacturing industry. In research on inventory management in multi-node and multi-product mode supply chains, Sultana et al. [28] used multi-agent reinforcement learning (MARL) to solve inventory management problems for multiple products and nodes. However, their environmental modeling is too simplistic, particularly in modeling the central warehouse. Choosing a method similar to the \((r, S)\)-policy of completely replenishing the warehouse will result in high additional inventory costs. Demizu et al. [29] used the RL-based method to solved the new product inventory management problem. Jullien et al. [30] researched the waste reduction via RL in supply chain inventory management. Wang et al. [20] modeled the supply chain management environment in the aviation manufacturing industry and used RL to solve inventory management problems for multiple materials under a single node. However, their assumptions about material demand are too simplistic and do not consider the dependencies between different types of material demand in different production processes.
While there have been some methods developed for managing inventory in simple supply chain environments, there are still several areas that need to be improved for large-scale manufacturing under the main manufacturer-supplier model in the field of supply chain inventory management:

- Demand modeling is too simple and fails to reflect the correlation between materials and process flow and the dependency relationship between different process flows;
- Different participants play different roles in the supply-demand relationship and have different goals and expectations for inventory management. Single-agent RL algorithms or homogenous MARL algorithms have limited further efficiency improvement;
- As the number of supply chain participants and the types of materials increases, the state space and action space of the environment grow explosively, requiring more efficient exploration of the policy space;
- With the increase in the number of supply chain participants, the requirement for internal collaborative cooperation within the supply chain will also increase. More effective methods are needed to prevent coupling when supply chain participants update their policies.

In order to solve the issues mentioned above, we proposed a supply chain inventory management method for civil aircraft manufacturing based on MARL. The main contributions of this paper can be summarized as follows:

- We formalized the supply chain inventory management problem of main manufacturer-supplier as a partially observable Markov decision process model.
- Based on the model, we proposed a MARL method for supply chain inventory management, in which the dual-policy and information transmission mechanism was designed to help the supply chain participant improve the global information utilization efficiency of the supply chain and the coordination efficiency with other participants.
- Based on the environment abstracted from actual civil aircraft manufacturing, we train and evaluate the proposed method. Experimental results show that our method has about 45% improvement compared with current RL-based supply chain inventory management methods.

2. POMDP Modeling

This section briefly describes the POMDP modeling of the supply chain inventory management problem of civil aircraft manufacturing addressed in this paper.

2.1. General Assumptions

To formalize the supply chain in the main manufacturer-supplier model, we make certain assumptions and abstractions about the actual supply chain, the main assumptions being:

- All participants in the supply chain make rational decisions and are not affected by uncontrollable factors such as politics;
- The time granularity of the model is weekly;
- The time required for material storage, withdrawal, and inventory count is not considered to emphasize the critical points of the issue;
- The circulation cost between warehouses within a particular party is not considered to highlight the interaction between the parties involved in the supply chain;
- The supplier selection problem is not considered based on the actual production situation, which means a unique supplier provides each material;
- The material requirements of the inventory department proposed by the production line are based on the types of processes rather than the entire aircraft;
- The start time nodes for preparing materials for each flight within the annual plan are evenly distributed throughout the year.
2.2. Demand Modeling

Unlike most previous research on supply chain inventory management, the research background of this paper is the inventory management problem of large-scale manufacturing supply chains, where the material demand has interdependent relationships between previous and subsequent processes. Taking a COMAC large passenger aircraft as an example, the overall assembly process involves process dependencies, as shown in Figure 1. The production of the COMAC civil aircraft is completed through a series of sequential and parallel processes, so a certain degree of abstraction of the actual problem is carried out in the environment. The process of the main manufacturer is a combination of serial and parallel structures, as shown in Figure 2.

![Figure 1](image1.png)

**Figure 1.** The structure of the demand dependency model of the civil aircraft production in COMAC.

![Figure 2](image2.png)

**Figure 2.** Example of the structure of the demand dependency model of main manufacturer.

In actual production processes, each step involves many different types of materials. These materials are typically divided into three categories based on their purpose: structural components, system components, and standard parts. Structural components refer to parts that comprise the main load-bearing components of the aircraft, such as the nose, fuselage, wings, and tail. System components refer to the components of the onboard systems, such as power systems, control systems, fuel systems, and hydraulic systems. Standard parts refer to standardized general-purpose parts used in production and maintenance, such as bolts and nuts. There are significant differences in the required quantity of materials between structural components, system components, and standard parts in each process. Therefore, in our study, we selected a representative structural component, system component, and standard part for each process. In order to reflect the widespread use of standard parts in actual production processes, our model uses only two types of standard parts for all processes. For more detailed information on the environmental setting, please refer to the relevant content in the Experimental Results section.

Considering factors such as the manufacturing process defects of suppliers and operator errors of the main manufacturer, materials may need to be reworked or scrapped. To prevent delays in production schedules caused by such shortages, we consider the demand
for materials at each process as a value that fluctuates up and down with a certain level of randomness. Therefore, it is assumed that the demand for material $v$ by process $i$ follows a normal distribution $N(\mu_{uv}, \sigma_{uv}^2)$ with mean $\mu_{uv}$ and variance $\sigma_{uv}^2$.

For the main manufacturer, we follow the annual planning mechanism in actual production and set the number of production tasks in the environment. The start times of these production tasks are evenly distributed throughout the annual production cycle. After the start of the first process of each production task, the logistics demand of that process is directly sent to the main manufacturer. Each process has a maximum lead time for materials. The main manufacturer will be punished for stock-outs if it exceeds the maximum lead time to ensure the on-time progress of the production plan. After the main manufacturer has completed the preparation of the materials required for a certain process, all the following processes will be added.

For the supplier, its material demand depends on the actions of the main manufacturer. The main manufacturer also has a maximum waiting period for the supplier’s stock preparation, and the supplier will be punished for stock-outs if it exceeds the maximum waiting period.

2.3. POMDP Specification

For the multi-agent sequential decision problem in which the agents can only perceive partial information about the environment, the POMDP is often used for formal description. It is usually formally described by the tuple $< N, S, T, O, R, \gamma >$. The information represented by each symbol is as follows:

- $N$: represents the set of all agents involved in the problem, and $|N|$ is the number of agents.
- $S$: represents the state space of global states in the problem. $s_t \in S$ represents the global state at time $t$. It is a vector composed of the observation information $o_t^i$ of each agent, $s_t = [o_t^1, o_t^2, \ldots, o_t^{|N|}]$.
- $T$: represents the joint action space of agents in the problem, which is composed of the action spaces of each agent, $A = A^1 \times A^2 \times \cdots \times A^{|N|}$. Considering that there are different demand levels for different materials, we use the action demand to represent the ratio of the production capacity and inventory level of the agent when considering the actions of the agents in the model. Therefore, for the action space $A^i$ of agent $i$, assuming $M^i$ is the number of types of materials involved in the inventory of the agent, the action space $A^i = [0, 1]^{M^i}$. At time $t$, the action of agent $i$ is $a_t^i \in A^i$. It seems logical to use continuous variables directly as the actions, but using such a continuous action space in actual training is inconvenient. On one hand, it makes the action space size too large to exploration. On the other hand, due to the legality constraint of actions, the sampling of actions over the interval will be trimmed, causing the non-differentiable spike impact on the two endpoints of the feasible action domain. Therefore, the action space is discretized as $\{0, 0.01, 0.015, 0.02, 0.04, 0.05, 0.1, 0.2, 0.4, 1\}$.
- $O$: represents the joint observation information obtained by the agents in the environment. For agents, there are two types of reward information they can receive: individual reward and team reward. The individual reward is the reward obtained by the agent based on its local observation information, representing the cost of the agent itself. The team reward is the reward obtained by all agents based on the global state, representing the total cost of the supply chain. Minimizing the total cost of the supply chain is the common optimization target of all agents. The team reward is the sum of individual reward of all agents in the supply chain $r_t = \sum_{i=1}^{|N|} r_t^i$. The individual reward of agent $i$ is $r_t^i = r_t^i^{stock} + r_t^i^{stock} + r_t^i^{overflow}$ composed of the following three parts:

- **Inventory cost $r_t^i^{stock}$**: the cost incurred to maintain inventory, including storage costs, storage management costs, etc. At time $t$, the inventory ratio of mater-
The demand observation information represents the order demand information.

- Shortages cost \( r_{t}^{\text{lack},j} \) is the cost incurred due to the inability to meet the demand for materials that are due to insufficient inventory. At time \( t \), the proportion of material \( j \) in the unfulfilled due orders of agent \( i \) is \( lck_{t}^{ij} \), and the stock shortage cost of the agent is \( r_{t}^{\text{lack},j} = \frac{\sum_{i=1}^{M} lck_{t}^{ij}}{M} \).

- Overflow cost \( r_{t}^{\text{overflow},j} \) is the additional storage cost incurred due to excess inventory. At time \( t \), the proportion of material \( j \) that exceeds the upper limit of agent \( i \)’s inventory capacity is \( ovf_{t}^{ij} \), and the overflow cost of the agent is \( r_{t}^{\text{overflow},i} = \frac{\sum_{j=1}^{M} ovf_{t}^{ij}}{M} \).

- \( O \): represents the joint observation space of agents in the problem, which is composed of the observation spaces of each agent, \( O = O_{1} \times O_{2} \times \cdots \times O_{|\mathcal{N}|} \). At time \( t \), the observation information of agent \( i \) is \( o_{t}^{i} \in O_{i} \). The observation information of each agent includes its inventory information and demand information. In addition, the observation information for the main manufacturer agent includes its logistics information, and for the supplier agent includes its production information.

- The demand observation information represents the proportion of various materials of the agent. At time \( t \), for agent \( i \) it is: \( inv_{t}^{i} = [inv_{t}^{i,1}, inv_{t}^{i,2}, \ldots, inv_{t}^{i,M}] \).

- The demand observation information represents the order demand information for various materials of the agent. At time \( t \), the material \( j \) demand observation information of agent \( i \) is \( ord_{t}^{ij} = [ord_{t}^{ij,0}, ord_{t}^{ij,1}, \ldots, ord_{t}^{ij,w}] \), where \( w \) is the maximum waiting time for the material \( j \) and \( ord_{t}^{ij,k} \) is the ratio of the absolute number of material \( j \) demand with remaining order duration time \( k \) to the maximum inventory limit of material \( j \) for agent \( i \) at time \( t \).

- The logistics observation information represents the transportation information for various materials of the main manufacturer. At time \( t \), the logistics observation information of material \( j \) of the main manufacturer is \( trn_{t}^{i} = [trn_{t}^{i,0}, trn_{t}^{i,1}, \ldots, trn_{t}^{i,j}] \), where \( l \) is the maximum transportation time for the material and \( trn_{t}^{i,j} \) is the ratio of the absolute number of material \( j \) transportation with remaining transportation time \( k \) to the maximum inventory limit of material \( j \) for the main manufacturer at time \( t \).

- The production observation information represents the production information for the corresponding material of the supplier. At time \( t \), the production observation information of material \( j \) of supplier \( i \) is \( pdt_{t}^{ij} = [pdt_{t}^{ij,0}, pdt_{t}^{ij,1}, \ldots, pdt_{t}^{ij,u}] \), where \( u \) is the maximum production time for the material and \( pdt_{t}^{ij,k} \) is the ratio of the absolute number of material \( j \) production with remaining production time \( k \) to the maximum inventory limit of material \( j \) for supplier \( i \) at time \( t \).

- \( \gamma \in [0, 1] \) is the discount factor. It represents the importance of future rewards compared to immediate rewards for the agents.

3. Proposed Supply Chain Inventory Management Method

In the prior section, we modeled the supply chain inventory management in civil aircraft manufacturing. In this section, we elaborate on the motivation of this paper and formally describe the problem studied in this paper. Starting from motivation, we introduce our MARL method based on the dual policy model with the information transmission mechanism based on Multi-Agent Proximal Policy Optimization (MAPPO) [31] and its corresponding theoretical derivation. Compared to the baseline method, MAPPO, our proposed method introduces a dual policy mechanism and an information transmission.
mechanism to help the supply chain participant improve the global information utilization efficiency of the supply chain and the coordination efficiency with other participants.

3.1. Method Overview

There are few common MARL environments for inventory management problems based on the main manufacturer-supplier model in the supply chain. Therefore, we first formally describe the inventory management problem under this model to construct a RL environment for training. As different participants in the supply chain play different roles and require different types and quantities of materials for inventory management, some differences exist in inventory management methods and goals. Suppose a single-agent RL algorithm [18] or homogeneous MARL algorithm [19] is used to address this problem. In that case, the model’s efficiency will be limited to some extent. Therefore, it is necessary to distinguish between the main manufacturer and the supplier when modeling and use heterogeneous agents for modeling to be as close as possible to the actual production environment.

As shown in the Figure 3, the policy network and value network of the agent take the data obtained from the agent’s interaction with the environment as input. The input data first undergoes feature extraction using a Multi-Layer Perceptron (MLP), and the extracted features are then inputted into a Recurrent Neural Network (RNN). The output of the RNN is then fed into another MLP to obtain the final actions from the policy network or the final value from the value network. The generated actions and values are then saved in the data buffer. After collecting data for the entire trajectory through interactions, the policy network and value network of the agent will be updated with the data stored in the data buffer. The specific details and configurations related to the network structure will be elaborated on in the Section 4.

Figure 3. The structure of the MARL method based on the dual policy model with the information transmission mechanism.

As mentioned earlier, to be close to the real production environment, we made certain assumptions about the process flow and types of materials based on the actual production process. Due to a large number of materials and participants in the supply chain in the environment, on the one hand, it will result in a large state and action space of the environment, which is not conducive to policy exploration. On the other hand, when
different agents update their policies, there will be significant coupling effects on the joint policy of the supply chain, and the improvement of the joint policy may not be apparent.

In RL problems with multiple agents cooperating in tasks such as supply chain inventory management, the agents in the environment explore and update their policies to maximize the team reward. In actual training, on the one hand, the state space and action space of the environment explode with the increase in supply chain participants and types of materials, leading to low efficiency in policy exploration. On the other hand, as each supply chain participant shares the same team reward, they lack sufficient information to judge the impact of their policy updates on the joint policy of the supply chain [32,33]. Therefore, we introduce a dual-policy mechanism consisting of independent and team policies that maximize individual and team rewards, respectively. By comparing the two policies, supply chain participants can understand the impact of their policy updates on the joint policy of the team. In the initial training stage, independent policies are used for exploration as much as possible. After the policies are relatively stable, the agents gradually switch to relying on team policies for decision-making using hyperparameters, which can significantly improve the efficiency of the agents’ exploration.

Furthermore, existing methods lack information transmission between the main manufacturer and the supplier when updating their policies. It will affect the exploration of the global optimal inventory management joint policy for the supply chain. The supply chain participants only use their local information to update their policies, ignoring the influence of others participants’ policies. It leads to a coupling phenomenon of the overall joint policy of the cooperative participants during policy updates. As a result, the monotonicity of the performance of the joint policy cannot be guaranteed during the participants’ policy updates. Therefore, we introduce the information transmission mechanism in the model. The influence of the participants’ updates on the individual joint policy and the team joint policy is recorded separately. The information recorded is passed among the participants by the update order. It ensures the monotonicity of the joint policy updates when the participants update their policies.

When the training of the method is completed, for each participating agent in the supply chain, the relevant information $o_i^t$ can be used as input to obtain recommended procurement amounts (for the main manufacturer) or production amounts (for the supplier) $a_i^t$ for the various materials involved by that supply chain participant. The practitioners of supply chain can improve the efficiency of material inventory management via this method.

### 3.2. Dual Policy Model with Information Transmission

In the dual policy mechanism of the method, each agent $i$ maintains a pair of individual policy $\pi^i$ and team policy $\hat{\pi}^i$ separately. The individual policy $\pi^i$ is used for direct interaction with the environment, and the team policy $\hat{\pi}^i$ is learned from the information obtained from the environment obtained from the environment.

According to the PPO algorithm [26], the update of the individual policy $\pi^i$ is subject to update step size constraints. If $\eta_i^t(\theta^i)$ is the update step size, the individual policy $\pi^i$ updates with $J^{PPO}(\theta^i)$ as the update target according to Equation (1):

$$J^{PPO}(\theta^i) = \mathbb{E}[\text{clip}(\eta_i^t(\theta^i), 1 \pm \epsilon) A_i^t]$$

$$\eta_i^t(\theta^i) = \frac{\pi_{\theta^i}(a_i^t|o_i^t)}{\pi_{\theta_{old}^i}(a_i^t|o_i^t)}$$
where $\epsilon$ is a hyperparameter that limits the update step size, and $A^i_t$ is the advantage function of agent $i$ estimated by the GAE algorithm \cite{28} as shown in Equation (2):

$$
A^i_t = \sum_{h=0}^{h} (\gamma^h \delta^i_t)
$$

$$
\delta^i_t = r^i_t + \gamma V^i_{\hat{\pi}}(s_{t+1}) - V^i_{\pi}(s_t)
$$

where $\delta^i_t$ is the TD error of agent $i$ at time $t$ for the trajectory of length $h$.

In our model, the individual reward $r^i_t$ and the team reward $r_t$ are made use of by each agent to learn the individual policy $\pi^i$ and the team policy $\hat{\pi}$. To ensure that the individual policy $\pi^i$ does not deviate too much from the goal of maximizing team profits during the exploration process, we define the similarity between the individual and team policies as $\sigma^i(\theta^i)$, at observation $o^i_t$ and action $a^i_t$. The policy update also subject to the constraint $J^{IDV-TEAM}(\theta^i)$ as Equation (3):

$$
\sigma^i(\theta^i) = \frac{\pi_{\hat{\pi}^i}(a^i_t | o^i_t)}{\pi_{\pi^i}(a^i_t | o^i_t)}
$$

$$
J^{IDV-TEAM}(\theta^i) = \mathbb{E}[clip(\sigma^i(\theta^i), 1 \pm \xi) A^i_t]
$$

where $\xi$ is a hyperparameter that limits the update step size.

Next, we need to combine the information got from Equations (1) and (3) for the update of individual policy. Since the individual policy $\pi^i$ is mainly used to assist exploration during the early stages of training, it should still be updated with the goal of maximizing group interests. The data used for its own training does not include information on team returns $r_t$, so it should be made close to the direction of the team policy $\hat{\pi}$ during training. We can get following conclusion from the definition of the similarity $\sigma^i(\theta^i)$ between the individual and team policies. When $\sigma^i(\theta^i) \leq 1$ and $A^i_t > 0$, it means that under the observation information $o^i_t$, the individual policy $\pi^i(a^i_t | o^i_t)$ is consistent with the team policy $\hat{\pi}(a^i_t | o^i_t)$, so when $\sigma^i_t \leq 1$, it should be ensured that:

$$
J(\theta^i) = \mathbb{E}[\|\sigma^i_t\|_{\leq 1} \max(J^{PPO}(\theta^i), J^{IDV-TEAM}(\theta^i))]
$$

Similarly, when $\sigma^i_t > 1$ it should be ensured that:

$$
J(\theta^i) = \mathbb{E}[\|\sigma^i_t\|_{> 1} \min(J^{PPO}(\theta^i), J^{IDV-TEAM}(\theta^i))]
$$

For the individual policy to learn the team policy information appropriately, we use the KL divergence with a hyperparameter $\alpha$ that gradually increases as the training progresses as a constraint. Therefore, the complete objective function for the individual policy $\pi^i$ is as follows:

$$
J(\theta^i) = \mathbb{E}[\|\sigma^i_t\|_{\leq 1} \max(J^{PPO}(\theta^i), J^{IDV-TEAM}(\theta^i))]
+ \mathbb{E}[\|\sigma^i_t\|_{> 1} \min(J^{PPO}(\theta^i), J^{IDV-TEAM}(\theta^i))]
- \alpha KL(\hat{\pi}, \pi^i)
$$

(6)

As the data used for learning is obtained by individual policy $\pi^i$, the importance sampling is used to update the team policy $\hat{\pi}$:

$$
\hat{\sigma}^i(\theta^i) = \frac{\pi_{\hat{\pi}^i}(a^i_t | o^i_t)}{\pi_{\pi^i}(a^i_t | o^i_t)}
$$

$$
J(\theta^i) = \mathbb{E}[clip(\hat{\sigma}^i(\theta^i), 1 \pm \xi) A^i_t - \beta KL(\pi^i, \hat{\pi})]
$$

(7)

where $\xi$ is a hyperparameter that limits the update step size, $\beta$ is a hyperparameter that gradually decreases with training, indicating that the influence of individual policy on the
latter stages of training decreases, and $\hat{A}_t$ is the estimated advantage function obtained through the GAE algorithm.

In the MAPPO algorithm, the joint policy is updated by using the local information of each agent to update the policy of each agent in the direction of each agent’s individual reward. In the cooperative problem, it is necessary for the agents to have a certain level of communication and coordination mechanism in order to prevent coupling of the joint policy of the participating parties during the update. So we introduce the information transmission mechanism in our model to solve this problem.

If $\{i_1, i_2, ..., i_m\}$ is a subset of the set of all agents $N$, where $i_k$ represents the $k$th element in the subset, and the complement of the subset is denoted by $-i_{1:m}$, then the state-action value function of the subset is:

$$Q_{\pi}^{i_{1:m}}(s, a^{i_{1:m}}) \triangleq \mathbb{E}_{a^{-i_{1:m}} \sim \pi^{-i_{1:m}}} [Q_{\pi}(s, a^{i_{1:m}}, a^{-i_{1:m}})]$$ (8)

For disjoint subsets $i_{1:m}$ and $j_{1:k}$, the multi-agent advantage function of subset $i_{1:m}$ can be calculated from the state-action value function as:

$$A_{\pi}^{i_{1:m}}(s, a^{j_{1:k}}, a^{i_{1:m}}) \triangleq Q_{\pi}^{j_{1:k}, i_{1:m}}(s, a^{j_{1:k}}, a^{i_{1:m}}) - Q_{\pi}^{j_{1:k}}(s, a^{j_{1:k}})$$ (9)

Therefore, if we specify an update order from $i_1$ to $i_m$ for the agents being updated, we can obtain:

$$A_{\pi}^{i_{1:m}}(s, a^{i_{1:m}}) \triangleq \sum_{j=1}^{m} A_{\pi}^{i_j}(s, a^{i_{1:j-1}}, a^{i_j})$$ (10)

The Equation (10) does not require any additional assumptions beyond specifying an update order for the agent’s update, and subsequent agents can further update based on the understanding of the influence of the previous agents’ update on the joint policy, ensuring monotonic update of the joint policy in a single update. Therefore, we maintain the information transmission factors of the individual joint policy and the team joint policy based on this. The factors record the update status of the joint policy to achieve monotonic update of the joint policy. In addition, we randomly reassign the update order of the agents at each update to eliminate the influence of a specific update order on policy update.

4. Experiment Results

In this section, we verify the efficiency advantage of our MARL method based on the dual policy model with the information transmission mechanism in the environment by compared with some existing methods.

4.1. Experiment Environment

The process flow in the experimental environment is shown in the Figure 2, and the material demand of the process flows involved in the experiment scenario setting is shown in the Table 2. The material demand shown in the Table 2 is derived from the abstraction and analysis of historical data obtained through our on-site investigation at COMAC Shanghai Aircraft Manufacturing Co., Ltd.

Due to the management of 22 materials in the scenario, the cooperation among totaling 23 agents, including 22 suppliers and one main manufacturer, is considered in the environment as mentioned in the section Model Assumptions. The parameters of the scenario used in the experiments are shown in Table 3. The information about the network in the model mentioned in our proposed method is shown in Table 4.
Table 2. The table of parameters about material demand.

<table>
<thead>
<tr>
<th>Assembling Process</th>
<th>Material Category</th>
<th>Demand Mean</th>
<th>Demand Std</th>
</tr>
</thead>
<tbody>
<tr>
<td>assembling the center wing</td>
<td>structural component</td>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>system component</td>
<td>150</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>standard component A</td>
<td>1500</td>
<td>50</td>
</tr>
<tr>
<td>assembling the center fuselage</td>
<td>structural component</td>
<td>50</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>system component</td>
<td>300</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>standard component B</td>
<td>2000</td>
<td>50</td>
</tr>
<tr>
<td>assembling the aft fuselage</td>
<td>structural component</td>
<td>50</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>system component</td>
<td>300</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>standard component A</td>
<td>2000</td>
<td>50</td>
</tr>
<tr>
<td>assembling the nose</td>
<td>structural component</td>
<td>50</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>system component</td>
<td>300</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>standard component B</td>
<td>2000</td>
<td>50</td>
</tr>
<tr>
<td>joining the fuselage</td>
<td>structural component</td>
<td>40</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>system component</td>
<td>500</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>standard component A</td>
<td>3000</td>
<td>50</td>
</tr>
<tr>
<td>assembling the tailfin</td>
<td>structural component</td>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>system component</td>
<td>150</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>standard component B</td>
<td>1500</td>
<td>50</td>
</tr>
<tr>
<td>installing the wing stabilizer</td>
<td>structural component</td>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>system component</td>
<td>150</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>standard component A</td>
<td>1500</td>
<td>50</td>
</tr>
<tr>
<td>joining the entire aircraft</td>
<td>structural component</td>
<td>40</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>system component</td>
<td>100</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>standard component B</td>
<td>2000</td>
<td>50</td>
</tr>
<tr>
<td>installing system components</td>
<td>structural component</td>
<td>25</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>system component</td>
<td>200</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>standard component A</td>
<td>1000</td>
<td>50</td>
</tr>
<tr>
<td>installing engines</td>
<td>structural component</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>system component</td>
<td>150</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>standard component B</td>
<td>2000</td>
<td>50</td>
</tr>
</tbody>
</table>

Table 3. The table of parameters about the scenario.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>number of agents</td>
<td>23</td>
</tr>
<tr>
<td>length of trajectories</td>
<td>400</td>
</tr>
<tr>
<td>number of PPO epoch</td>
<td>10</td>
</tr>
<tr>
<td>learning rate</td>
<td>$7 \times 10^{-4}$</td>
</tr>
<tr>
<td>discount factor $\gamma$</td>
<td>0.99</td>
</tr>
<tr>
<td>clip $\epsilon$</td>
<td>0.2</td>
</tr>
<tr>
<td>clip $\zeta$</td>
<td>0.2</td>
</tr>
<tr>
<td>clip $\xi$</td>
<td>0.2</td>
</tr>
<tr>
<td>GAE $\lambda$</td>
<td>0.95</td>
</tr>
<tr>
<td>KL coefficient $\alpha$ in $J(\theta)$</td>
<td>$0.0 \rightarrow 1.0$</td>
</tr>
<tr>
<td>KL coefficient $\beta$ in $\hat{J}(\hat{\theta})$</td>
<td>$1.0 \rightarrow 0.0$</td>
</tr>
</tbody>
</table>
### Table 4. The table of information about the network of agent $i$.

<table>
<thead>
<tr>
<th>Sub-Network</th>
<th>Parameters</th>
<th>Critic Network Setting</th>
<th>Policy Network Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLP (input)</td>
<td>number of the hidden layers</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>hidden size</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>output feature size</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>RNN</td>
<td>number of the recurrent layers</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>input feature size</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>output feature size</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td>MLP (output)</td>
<td>number of the hidden layers</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>hidden size</td>
<td>64</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>output feature size</td>
<td>1</td>
<td>size of $a_i$</td>
</tr>
</tbody>
</table>

### 4.2. Compared Methods

In the scenario described above, to compare with existing supply chain inventory management research, we select several methods that have performed well in recent studies by researchers as comparisons. The selected methods are as follows:

**SRL**: The method that treats the problem as a single-agent reinforcement learning problem [18].

**MAPPO**: There are some research about supply chain inventory management based on MARL [28], and our proposed method is modified from the MAPPO algorithm. So the MAPPO algorithm is selected for comparison.

**DualPolicy-w/oIT (ours)**: Our method, but without the information transmission mechanism as an ablation experiment for comparison.

**DualPolicy-IT (ours)**: Our MARL method based on the dual policy model with the information transmission mechanism.

The experimental results are obtained by training the agents in the environment with different algorithms. During the training, 10 environment parallel samples are used, and the length of each episode is 1000. For each algorithm or control experiment, 10 training was performed, and the solid line in the results represents the average of ten trainings, and the shadow represents the maximum and minimum values of 10 trainings.

### 4.3. Results and Analysis

We compared the experimental results of the four methods mentioned above in the environments described as Tables 2 and 3. The results are shown in the Figures 4–17. The Figure 4 represents the team reward of every episode, which is a macro-performance evaluation index. It means the total cost of the supply chain in the environment. The Figures 6–17 represent the average step inventory costs of the main manufacturer and the 22 suppliers, which are micro-performance evaluation indexes in the environment. For the convenience of the method policy updates and observations, we uniformly processed the positive and negative of all evaluation indexes: a higher numerical value indicates better performance of the method in the corresponding evaluation index. The comparison of the result curve in the same result figure concerning the variation of episodes is also a contrasting reference for the impact of the number of iterations on the experiment results.

### 4.3.1. The Macro-Performance Evaluation

Figure 4 is the macro-performance evaluation index, showing the variation of the overall supply chain team reward during the training. The methods, **DualPolicy-IT (ours)** and **DualPolicy-w/oIT (ours)**, gradually converge to approximately $-210$ and $-300$ levels, respectively, around episode 400. On the other hand, the methods **MAPPO** and **SRL** start to converge to relatively stable results around episode 800, at levels of approximately...
−380 and −420, respectively. The DualPolicy-IT(ours) method achieves approximately a 45% performance improvement compared to MAPPO.

Figure 4. The comparison of four methods described about episode team reward.

In our proposed method, to minimize the impact of the learning rate on method performance, we used a trick called learning rate decay. In order to explore the influence of the learning rate on the method, we conducted a set of control experiments, and the experiment results are shown in Figure 5. The control group used $7 \times 10^{-4}$ as the initial learning rate, and the initial learning rates of the experimental group were $7 \times 10^{-2}$, $7 \times 10^{-3}$, and $7 \times 10^{-5}$. The experiment results show that the large or small learning rate has a certain degree of influence on the performance of the method. When the learning rate is too large, such as $7 \times 10^{-2}$, the learning rate will cause the span of each update of the method to be too large, failing to converge to the optimum, which will degrade the performance of the method. However, we set learning rate decay. It can be clearly seen that as the training progresses, the impact of excessive learning rate has been alleviated to a certain extent, and finally converged to a final level slightly lower than the results of the control group at about episode 950. When the learning rate is too small, such as $7 \times 10^{-5}$, the learning rate will result in a small update range, making the performance of the method trapped in a local optimum. The $7 \times 10^{-5}$ only finally converged to about −1500 at episode 600.

Figure 5. The comparison of four initial learning rates.

4.3.2. The Micro-Performance Evaluation

Figure 6 shows the variation of average step inventory costs of the main manufacturer during the training process, which serves as a micro-performance evaluation index. Compared to the MAPPO method, the DualPolicy-IT (ours) and DualPolicy-w/oIT (ours)
methods, due to the dual policy mechanism, have already converged to a higher level around episode 400. However, the MAPPO method gradually stabilizes after episode 600. Compared to the DualPolicy-w/oIT (ours) method, the DualPolicy-IT (ours) method, thanks to the information transmission mechanism, achieves approximately a 25% improvement in the final convergence results. Due to the instability caused by treating the multi-agent problem as a single-agent problem, the SRL method is far inferior to the other three methods in terms of convergence speed and results.

Figure 6. The comparison of four methods described about inventory cost of the main manufacturer.

Figure 7 shows the variation of average step inventory costs of the two standard component suppliers during the training process, which serves as a micro-performance evaluation index. In this set of experiment result figures, compared to MAPPO and SRL, the DualPolicy-IT (ours) and DualPolicy-w/oIT (ours) methods, with the dual policy mechanism, can effectively explore the large state space and action space in the real-world environment, helping the model converge faster.

Figures 8, 12 and 15 show the variation of average step inventory costs of the structural component supplier and the system component supplier of the central wing, the fuselage and the entire aircraft joining during training, serving as the micro-performance evaluation indexes. In this set of experiment result figures, it can be observed that due to treating the multi-agent problem as a single-agent problem, the SRL method leads to significant fluctuations in the results.

(a) Inventory cost of the standard component A supplier. (b) Inventory cost of the standard component B supplier.

Figure 7. The comparison of four methods described about inventory cost of the standard component suppliers.
As the results are shown in Figure 8, 11, 14 and 15, the method DualPolicy-IT (ours) also displays some fluctuations. This may be due to small policy changes leading to significant variations when mapped to the evaluation index space. However, from the experimental results, we can observe that compared to other methods, DualPolicy-IT (ours) tends to better handle these fluctuations to avoid getting trapped in the local optimum and converges to superior results. Based on the comparison of results with different episode sizes in the experiments, we recommend setting the episode size to at least 700 or above to ensure the stability of the method training results.

Figure 8. The comparison of four methods described about inventory cost of the center wing suppliers.

Figure 9. The comparison of four methods described about inventory cost of the center fuselage suppliers.
(a) Inventory cost of the center fuselage structural component supplier  
(b) Inventory cost of the center fuselage system component supplier

Figure 8. The comparison of four methods described about inventory cost of the center fuselage suppliers.

(a) Inventory cost of the aft fuselage structural component supplier  
(b) Inventory cost of the aft fuselage system component supplier

Figure 9. The comparison of four methods described about inventory cost of the aft fuselage suppliers.

(a) Inventory cost of the structural component supplier.  
(b) Inventory cost of the system component supplier.

Figure 10. The comparison of four methods described about inventory cost of the aft fuselage suppliers.

Figures 9, 10, 11, 14, 16 and 17 show the variation of average step inventory costs of the structural component supplier and the system component supplier of the center fuselage, the aft fuselage, the nose, the wing stabilizer, the system installing and the engine during training, serving as the micro-performance evaluation indexes. In this set of experiment result figures, compared to MAPPO and SRL, the DualPolicy-IT (ours) and DualPolicy-w/oIT (ours) methods, with the dual policy mechanism, can effectively explore the large state space and action space in the real-world environment, helping the model converge faster.

Figure 13 also shows the variation of average step inventory costs of the structural component supplier and the system component supplier of the tailfin during training, serving as the micro-performance evaluation index. In this set of experiment result figures, compared to the DualPolicy-w/oIT (ours) method, the DualPolicy-IT (ours) method, thanks to the information transmission mechanism, achieves evident performance improvement in the convergence level results.

(a) Inventory cost of the structural component supplier.  
(b) Inventory cost of the system component supplier.

Figure 11. The comparison of four methods described about inventory cost of the nose suppliers.
(a) Inventory cost of the structural component supplier. (b) Inventory cost of the system component supplier.

Figure 10. The comparison of four methods described about inventory cost of the nose suppliers.

Through these evaluation indexes, we can clearly see that the SRL method is far inferior to the other three methods in terms of convergence speed and results. It is due to the instability caused by treating the multi-agent problem as a single-agent problem and using single-agent reinforcement learning to solve it.

Since the MARL method can alleviate the instability of the environment when solving the multi-agent problem, the MAPPO method before improvement has a significant improvement in the convergence speed and fluctuation range compared with the SRL method in Figures 4 and 6. The participants of the supply chain in this method lack the effective utilization of global information, so they cannot further learn which actions are conducive to maximizing the team interests of the whole supply chain. The agents mainly update the policy with the local information in this method. Therefore, as shown in Figures 9b and 11a, about the MAPPO there are still large fluctuations, even in the middle of training.

(a) Inventory cost of the structural component supplier. (b) Inventory cost of the system component supplier.

Figure 11. The comparison of four methods described about inventory cost of the fuselage suppliers.

Figure 12. The comparison of four methods described about inventory cost of the tailfin suppliers.

Figure 13. The comparison of four methods described about inventory cost of the tailfin suppliers.
Compared with the existing two types of methods, the method proposed in this paper has obvious advantages in terms of convergence results and convergence speed. The ablation experiment we designed introduces a dual policy mechanism based on the MAPPO method. The DualPolicy-w/off (ours) method can effectively use the team information of supply chain to guide individual participants to update their policies. Therefore, through the comparison in Figures 4, 6, 12a and 13a, we can clearly observe that the comprehensive utilization of individual and team information guides the direction of policy update, effectively avoiding the fluctuation of the performance of the method during convergence. By comparing the results in the ablation experiment with or without the information transmission mechanism, the information transmission mechanism better coordinates the cooperation between the participants in the supply chain when updating the joint policy. Therefore, DualPolicy-IT (ours) has obvious advantages in terms of convergence speed compared with the other three methods. In the term of episode team reward, our model has efficiency improvement compared with existing methods, such as in Figure 14b.

Figure 14. The comparison of four methods described about inventory cost of the wing stabilizer suppliers.

(a) Inventory cost of the structural component supplier. (b) Inventory cost of the system component supplier.

Figure 15. The comparison of four methods described about inventory cost of the entire aircraft joining suppliers.

(a) Inventory cost of the structural component supplier. (b) Inventory cost of the system component supplier.
(b) Inventory cost of the entire aircraft joining structural component supplier

Figure 15. The comparison of four methods described about inventory cost of the system installing suppliers.

(a) Inventory cost of the structural component supplier.  (b) Inventory cost of the system component supplier.

Figure 16. The comparison of four methods described about inventory cost of the system installing suppliers.

(a) Inventory cost of the structural component supplier.  (b) Inventory cost of the system component supplier.

Figure 17. The comparison of four methods described about inventory cost of the engine suppliers.

5. Conclusions and Future Work

5.1. Conclusions

In this paper, we proposed a solution to supply chain inventory management for civil aircraft manufacturing. We formalized the problem as a POMDP model, where the assumption of material demand based on the actual production process can effectively reflect the correlation between materials and processes, as well as the temporal dependencies among materials in different processes. We proposed a MARL method with non-sharing policy for civil aircraft manufacturing inventory management, allowing participants in the supply chain to independently learn their own policies based on their local information. Our method with the dual-policy and information transmission mechanism was designed to improve the global information utilization efficiency of the supply chain and coordination among the participants. We evaluated our method with a scenario abstracted from the actual civil aircraft manufacturing. The experiment results show that our model has about 45% performance improvement compared with existing RL-based methods, which means our method can effectively manage the inventory of both the main manufacturer and the suppliers in the civil aircraft manufacturing supply chain.

5.2. Future Work

In addition to the factors we have considered in our current work, another characteristic of supply chain inventory management is that the demand for certain types of
materials exhibits temporal dependencies, while for others, the demand shows significant asynchrony. If we can model and learn this feature of material demand, it would be beneficial for supply chain inventory management. For future work, on the basis of the proposed method, we plan to further improve the performance of the method by learning the causal relationship between the production processes and the materials via the attention mechanism to promote the cooperation of the supply chain participants in the next stage.

**Author Contributions:** Conceptualization, H.L. and R.L.; Methodology, M.P. and D.Z.; Validation, M.P. and D.Z.; Writing—original draft, M.P. and D.Z.; Writing—review & editing, M.P., D.Z., H.L., and R.L. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research is funded by National Key R&D Program of China (No. 2021YFB3301900).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data is contained within the article.

**Conflicts of Interest:** The authors declare no conflict of interest.

**References**


5. Cantini, A.; Peron, M.; De Carlo, F.; Sgarbossa, F. A decision support system for configuring spare parts supply chains considering different manufacturing technologies. *Int. J. Prod. Res.* 2022, 2, 2–1. [CrossRef]


12. Cuartas, C.; Aguilar, J. Hybrid algorithm based on reinforcement learning for smart inventory management. *J. Intell. Manuf.* 2023, 34, 123–149. [CrossRef]


**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.