**Review Suggestions:**

Not understood:

* The warehouses have a seasonal demand, which is usually unadapted to non-recurrent networks.
* action space discretization : why not, the environment would also work with a continuous action space (order from 0 to maxcapacity from the factory) *Don’t we have this?*

less important:

* The warehouses have a seasonal demand, which is usually unadapted to non-recurrent networks.
* An improvement could be to have different means and variances for the demands of the warehouses *We kind of have this, at least changing demands*
* - Have the cost of production increase with the number of goods produced (like in macroeconomy models)?
* I think it could be interesting to test the agent on more than one seasonal cycle.
* Maybe a bit too technical at the beginning.

Important:

* Regarding the results on the second situation, it seems that the s-q policy and SARSA won't be able to learn, but maybe the REINFORCE can perform better if trained longer
* more readable results (labels and titles on graphs, more synthetic printing in the tesing file)
* deep Q network
* Nonetheless, we can observe that the convergence only occurs after 6000 training episodes of one entire seasonal cycle and despite a low number of warehouses (only 3, with 1 factory). We can thus wonder if the agent would have the ability to tackle a real-life issue with many more constraints. *Maybe we could comment on the real world application in the result section*
* It could be interesting to change the model of demand to see if its form has an influence and if the order of efficiency between algorithms is still the same with different (but still plausible) demand functions models.
* But maybe that 15 features is too much and the function should be less guided (for Sarsa features)
* The part on related work is not very rich in references on works on similar issues, maybe you can find what have been done in this domain.
* You may detail a bit more the features engineering part to make this part clearer (why you take this, not this), maybe add a picture + there is no the list of features in the appendix.
* When you describe your implementation of the REINFORCE algorithm, you say that you tested different functions for phi. It may be really interesting to have more information on the results given by these different functions.
* ask myself why the REINFORCE algorithm for the first environment (1 warehouse) is much more instable than the others, you may comment on this. *Answer: A lot of exploration due to low learning rate*