International Journal of World Economic Research
ISSN: 3080-9886 (Print), ISSN: 3080-9894 (Online) | Vol. 4, No. 1, 2026

Robust Carbon-Budget Screening for Offline Joint Inventory and Freight Control Under Emission-Factor Shift

Adrian Stewart, Eva Collins *

School of Computer Science, College of Computing, Georgia Institute of Technology, USA

* **Corresponding author**: (Email: eva.collins49@gatech.edu)

Abstract: Offline inventory and freight policies are increasingly evaluated under finite carbon budgets, yet a predicted-emissions filter does not itself prevent over-budget realized emissions when vehicle factors are uncertain. Building on the action-level pathwise carbon screen of carbon-aware conservative Q-learning (CA-CQL), which guarantees compliance for a predicted emissions ledger, we extend the screen to an uncertainty-bound feasibility projection that preserves behavior-supported replenishment proposals. The study constructs synthetic fixed logs for a supplier, distribution center, and three retailers and evaluates policy-independent demand and emission-error streams. Five offline training datasets contain 24,000 transitions in total, and each policy is evaluated over 250 held-out trajectories in each uncertainty condition. Under a 4,300 kg CO2e cap and emission multipliers in [0.8, 1.2], unfiltered behavior violates the cap in 30.0% of episodes, the nominal CA-CQL screen in 5.2%, and a 1.20 robust screen in 0%, while demand fill declines from 89.64% to 89.50%. Under mild [0.9, 1.3] and severe [1.1, 1.5] shifts, the 1.20 screen has 2.0% and 18.8% observed violations; thus the conditional guarantee does not extend past its assumed bound. A 1.35 factor reduces severe-shift violations to 6.4%. Comparisons with behavior cloning and support-regularized offline Q-learning separate constraint feasibility from policy quality, while common-random-number evaluation and seed-level paired tests quantify uncertainty. The result is a reproducible, auditable robust-dispatch interface rather than a claim of end-to-end optimal inventory reinforcement learning.

Keywords: Offline reinforcement learning, carbon budget, inventory and transportation, robust optimization, action projection, emission factor uncertainty.

1. Introduction

Coordinating replenishment and freight changes both operating performance and environmental exposure. Multi-echelon inventory coordination [1] and inventory routing [2] establish the coupling between upstream stock, downstream fill, and delivery activity. For a finite transport-carbon allowance, this coupling becomes sequential: a dispatch may appear locally attractive but use an allowance that will not be available during subsequent shortages. Li, Bian, and Lin [3] introduced carbon-aware conservative Q-learning (CA-CQL) with a behavior-support prior, a remaining-budget state, and a pathwise action screen. Their screen tests whether known immediate emissions fit within the remaining carbon budget and, under nonnegative deterministic emissions, guarantees compliance for the modeled ledger. Because it acts on individual dispatch actions, the screen can be attached to any proposal policy, which makes it a natural safety layer for offline inventory–freight control.

In a real dispatch system, emissions assessed before the vehicle moves can differ from activity-based emissions booked afterward. The difference may result from load variation, distance error, real-world vehicle factors, or accounting boundaries. These factors are not necessarily known when a discrete replenishment–transport action is selected. The key distinction is therefore between a policy that never exceeds its predicted budget and a policy that never exceeds the realized budget. The two are equivalent only when predictive accounting is exact or the dispatch filter accounts for uncertainty.

This paper studies a targeted extension: replace the point estimate in the CA-CQL pathwise screen with an upper envelope for realized freight emissions, then project a proposed joint action to the nearest physically and carbon-feasible candidate. The same projection is evaluated behind a fixed logged heuristic and a behavior-cloned proposal model. A conservative offline Q-network [4] is implemented as a diagnostic and compared with the behavior models; safety is reported together with service, since ceasing deliveries can satisfy any emissions cap while ruining inventory performance.

The contribution consists of a conditional guarantee under a declared factor uncertainty set; a concrete discrete-action recovery operator that retains the original shipment proposal where possible; and a paired 250-episode-per-condition evaluation with three levels of emission shift, two alternative carbon caps, and offline behavioral and value baselines. The simulator and operational factors are synthetic and fully disclosed. Our claim is about reliability of an implementable dispatch mechanism, not industrial validation or global policy optimality.

2. Related Research

2.1. Fixed-data Sequential Decision Learning

Offline policy learning is vulnerable to selecting actions poorly covered by its logging policy. Batch-constrained Q-learning [5] restricts such actions; BEAR [6] addresses errors propagated by bootstrapping; conservative Q-learning [4] penalizes unsupported action values. D4RL [7] demonstrates the importance of fixed-dataset evaluation, but its standard tasks do not contain synchronized retailer replenishment,

vehicle trip, and measured emissions records. Constrained Markov decision processes [8] and constrained policy optimization [9] formalize resource and safety limits, although a hard realized-emission budget calls for an action-level construction when immediate footprints can be bounded.

2.2. Inventory Routing and Freight Carbon

The pollution-routing problem [10] relates transport decisions to energy and emissions, and subsequent road-freight reviews [11] document key vehicle-related assumptions. Multi-period carbon-regulated inventory routing [12] embeds environmental regulation into operational planning. Direct-delivery stochastic routing [13], approximate dynamic programming [14], decomposition methods [15], and stochastic stockout/route-failure models [16] provide complementary solution strategies. Standard Markov decision modeling [17] explains why a remaining carbon balance belongs in the controller state even when the final budget screen is a deterministic operator.

Accounting consistency also matters. Corporate value-chain standards [18] and the Global Logistics Emissions Council framework [19] stress the definition of transport system boundaries. UK Government conversion factors [20] provide examples of public accounting coefficients, but the experiment here deliberately uses explicitly synthetic factors so that uncertainties and violation mechanisms are known. Adaptive conformal inference under shift [21] is pertinent to learning uncertainty margins from real measurements; carbon-aware distributionally robust procurement [22] confirms the broader relevance of uncertainty sets in supply-chain decisions. Neither is claimed to be implemented as part of the present factor-bound controller.

3. Joint Replenishment and Emissions Model

3.1. Finite-horizon Synthetic Network

The simulator consists of one upstream supplier, one distribution center (DC), and three retailers over H=40 decision periods. The joint discrete action is a=(u, q1, q2, q3, m). The upstream order u belongs to {0, 20, 40}; retailer allocations qi belong to {0, 5, 10}; and the transportation mode m is regular consolidated or express. Redundant express indices with no downstream allocation are eliminated, producing 159 actions. The DC and retailer beginning stocks are 45 and (12, 16, 20), respectively. Each period begins with the dispatch decision, updates the order and freight pipelines, receives arrivals according to the implementation schedule, and then realizes demand. Express allocations are available before the current period demand draw; regular allocations and supplier orders enter their respective pipeline arrays for later periods. These exact intra-period timing conventions, rather than an assertion about real-world travel days, define the reproducible instance.

Demand is a gamma–Poisson seasonal mixture with retailer means (3, 4, 5), a weekly sinusoidal factor, and gamma overdispersion. Demand not fulfilled immediately accumulates as negative net inventory and is charged backlog cost. The monetary objective combines inventory holding, shortage, upstream ordering, retailer-trip fixed charges, and a nominal freight-activity charge. The observation includes DC inventory, three retailer net inventories, two upstream pipeline slots, six retailer pipeline quantities, cumulative realized emissions normalized by the deployment cap, normalized period index, and two weekly phase encodings: 16 features in all. Physical action masking enforces current available DC stock, mode-specific loading capacities of 30 or 20 units, and an upstream inventory-position bound of 160.

3.2. Point-Estimated and Realized Freight Footprints

The modeled nominal emissions $\hat{e}(a)$ use a transparent trip formula. An upstream order contributes $42+0.30u$ units of kg CO_2e if $u>0$. A nonempty regular trip adds $66+1.4(q1+q2+q3)$. A nonempty express dispatch adds 36 for each served retailer plus twice the shipped units. These coefficients are not empirical vehicle factors; they were chosen as scenario parameters. The realized increment is $e_t(a)=\hat{e}(a)\xi_t$ where ξ_t is a factor error drawn independently of the chosen action and solely from the simulated evaluation episode's pre-generated daily error array.

At nominal test conditions ξ_t is uniform on [0.8, 1.2]. At mild shift it is uniform on [0.9, 1.3], and at severe shift on [1.1, 1.5]. The cumulative realized account is $C_{t+1}=C_t+e_t(a_t)$, with $C_0=0$, and the nominal 40-period ceiling is B=4,300 kg CO_2e. A screen using $\hat{e}(a)$ can be breached by a final trip that exceeds its remaining balance even though the pre-dispatch nominal estimate was acceptable. The experiment distinguishes C_t, the true simulated emissions ledger, from the nominal estimated amount used in feasibility scoring.

4. Robust Behavior-Preserving Screening

4.1. Adopted Nominal Screen and Robust Extension

The CA-CQL dispatch procedure of Li et al. [3] screens an action by $\hat{e}(a)\le B-C_t$. We implement this inequality unchanged as the nominal screen, applied on top of the physical feasibility mask of Section III-A. For a specified multiplicative bound $\bar{u}\ge 1$, we define $A_{\bar{u}}(s_t)=\{a \text{ in } A_{phys}(s_t): \bar{u}\,\hat{e}(a)\le B-C_t\}$. The principal robust method uses $\bar{u}=1.20$; a margin variant uses $\bar{u}=1.35$. This changes the carbon test, not the source of behavioral proposals or the environment dynamics. Updating C_t using realized emissions makes the comparison particularly stringent: even an up-to-date actual ledger cannot correct an underestimated footprint of the very next accepted dispatch.

Proposition 1 (conditional pathwise feasibility). Assume $\hat{e}(a)\ge 0$; $e_t(a)\ge 0$; and $e_t(a)\le \bar{u}\,\hat{e}(a)$ for every selected action and time. If $C_0=0$ and a_t is chosen from $A_{\bar{u}}(s_t)$ each period, then $C_H\le B$. Proof: from feasibility, $C_t+\bar{u}\,\hat{e}(a_t)\le B$. Because $e_t(a_t)\le \bar{u}\,\hat{e}(a_t)$, it follows that $C_{t+1}\le B$. Induction gives the stated result. The proof is deterministic and is conditional on an actual valid bound for every selected dispatch; it does not provide population-level probabilistic coverage when factor errors leave the set.

With [0.8, 1.2] errors, $\bar{u}=1.20$ satisfies the hypothesis. With [0.9, 1.3] and [1.1, 1.5] errors, it does not. The 1.35 variant restores the worst-case bound for the first shifted scenario only. A lower-than-100% observed violation rate in severe shift is therefore an empirical result, not a theorem. The zero-shipment action with zero upstream order is always physically feasible, hence the filtered candidate set is nonempty at every remaining budget.

4.2. Action Projection and Service Retention

Let a* be the candidate from the behavior or learned offline proposal model. If it already lies in A_ubar(s), it is executed without alteration. Otherwise, we choose a in A_ubar(s) minimizing d(a,a*)=0.3|u−u*|+sum_i|qi−qi*|+12 I(m≠m*)+0.03 ê(a). This explicit joint-action projection gives upstream quantity changes a smaller penalty than retailer allocations and discourages mode switching and unnecessary emissions. Because it selects the nearest feasible joint action rather than forcing zero dispatch, it can retain substantially more of the proposed replenishment activity, but it need not optimize long-run cost. All weights are common to the nominal, robust, and margin comparisons.

4.3. Offline Proposal Estimators

Five synthetic logged datasets are generated by a base-stock-like behavior controller, with 10% random feasible exploration. Each dataset has 120 episodes of 40 transitions, and contains (s, a, r, s_next, terminal, physical_mask, next_mask). A 96×96 SiLU behavior network learns masked categorical action likelihoods by cross-entropy. A separate 96×96 conservative Q-network uses a smooth-L1 Bellman error plus 0.15 times the feasible-action log-sum-exp penalty. Both fit only these saved transitions, using 300 updates, batch size 256, AdamW, discount 0.98, and target copies every 15 updates; no evaluation transition enters training.

The behavior cloning (BC) proposer chooses the highest-score physically feasible action. The support-CQL diagnostic proposes by a weighted combination of its learned value and behavior-prior log probability, with a nominal emission preference. Behavior, BC, and support-CQL are evaluated as fixed policies, while separate nominal and robust screens are attached to behavior and BC for direct comparisons. This isolates action-screen performance from whether a short-run offline Q fit discovers an economically effective dispatch policy.

5. Experimental Protocol

5.1. Independent Training Runs and Paired Environment Noise

Five independently generated synthetic offline datasets contain 24,000 total transitions, with a fresh BC and Q checkpoint trained for each dataset. For every policy, seed, uncertainty condition, and specified cap, 50 held-out evaluation episodes are run. The main test therefore has 5×50=250 episodes per policy per condition. The evaluation code generates a fixed 40-day exogenous demand array and a fixed 40-day error-multiplier array for each episode before any action is taken. Competing policies thus face identical daily uncertainty realizations even when their dispatch choices diverge. Model training and test sequences have disjoint episode seeds.

Metrics are undiscounted operating cost, realized kg CO2e, fraction of same-period retail demand fulfilled, number of nonempty retailer trips, and fraction of episodes exceeding the true carbon cap. We report the mean of five seed-level means and retain each episode in the raw CSV. We perform exact two-sided paired sign-flip randomization tests across the five seed-level contrasts; with five seeds the smallest possible two-sided p-value is 0.0625, so observed robust directions should not be recast as conventional $p<0.05$ findings. Individual test episodes are not treated as independent training replications.

5.2. Conditions and Component Controls

The main controls are unfiltered behavior, the original nominal screen, robust screening with 1.20 multiplier, and a more conservative 1.35 multiplier. We additionally report a consolidation-oriented heuristic, unscreened BC, BC with the 1.20 screen, and support-CQL without the upper-emission wrapper. Every method uses the same action and physical feasibility definitions. The three multiplier conditions test validity of the declared uncertainty set, while the screen-factor ablation isolates the conservatism. Cap sensitivity repeats nominal and robust behavior and BC without retraining at B=3,900 and 4,700 kg CO2e.

6. Results and Analysis

6.1. Main Results Across Uncertainty Regimes

Table I. Within-Bound Emission-Factor Errors

Policy	Cost	kg CO2e	Fill %	Viol. %
Logged heuristic	5,575	4,179	89.6	30.0
Nominal screen	5,597	4,136	89.5	5.2
Robust (1.20)	5,599	4,129	89.5	0.0
Margin (1.35)	5,599	4,125	89.5	0.0
Consolidation	7,095	4,153	75.1	24.8
Offline BC	28,946	4,000	50.3	15.6
BC + robust	28,953	3,972	50.3	0.0
Support-CQL	29,411	3,964	49.7	14.0

Table II. Mild Emission-Factor Shift

Policy	Cost	kg CO2e	Fill %	Viol. %
Logged heuristic	5,575	4,597	89.6	90.8
Nominal screen	6,219	4,271	87.2	19.6
Robust (1.20)	6,281	4,259	87.1	2.0
Margin (1.35)	6,350	4,249	86.9	0.0
Consolidation	7,095	4,569	75.1	89.2
Offline BC	28,939	4,405	50.6	64.0
BC + robust	29,276	4,192	50.3	3.6
Support-CQL	29,443	4,357	49.9	61.2

Table III. Severe Emission-Factor Shift

Policy	Cost	kg CO2e	Fill %	Viol. %
Logged heuristic	5,575	5,432	89.6	100.0
Nominal screen	12,353	4,295	76.8	44.4
Robust (1.20)	12,554	4,279	76.6	18.8
Margin (1.35)	12,721	4,270	76.4	6.4
Consolidation	7,095	5,399	75.1	100.0
Offline BC	29,004	5,204	50.9	96.8
BC + robust	32,554	4,278	47.9	21.2
Support-CQL	29,445	5,152	50.3	94.4

Under the within-bound condition, unfiltered behavior exceeds the true 4,300 kg limit in 30.0% of 250 episodes. The nominal CA-CQL screen of [3] already reduces this to 5.2%, removing most overruns; the remaining violations arise only from the gap between predicted and realized emissions, which robust 1.20 screening closes to 0.0%. Its fill rate is 89.50%, compared with 89.64% for unfiltered behavior, and the mean operating-cost difference is 24.7 synthetic units. In this regime compliance is obtained at relatively small mean service and cost penalties, rather than by switching off

logistics.

Once the true factors extend above 1.20, the mathematical implication ceases to apply. Under mild shift, the nominal and robust violation rates are 19.6% and 2.0%, while the 1.35 margin records 0.0%. The valid 1.35 upper factor closes the logical gap for this particular range. Under severe shift, nominal and 1.20 robust rates rise to 44.4% and 18.8%; even the 1.35 factor has 6.4% observed failures because it is itself below the 1.50 simulated maximum. The shift results highlight bound misspecification as the dominant threat to any hard-guarantee claim.

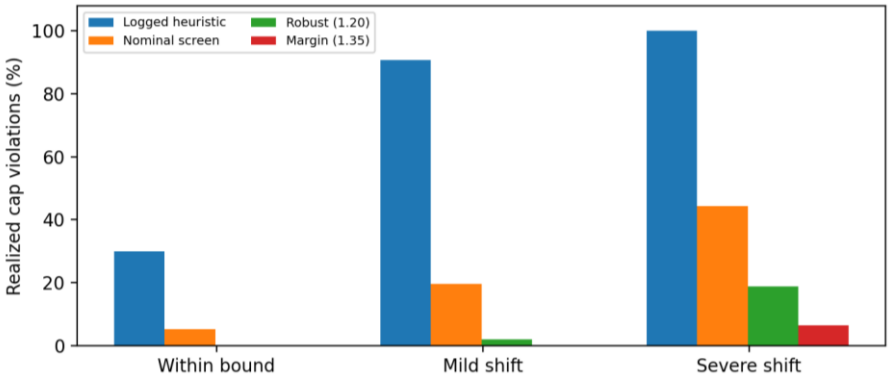


Figure 1. Violation rates under three emission-error regimes; each plotted value uses 250 evaluated trajectories.

6.2. Offline-learning Ablation and Operational Trade-Offs

The offline BC and support-CQL models are trained on the five fixed logs. In the nominal-range evaluation, unscreened BC fills 50.32% of demand and support-CQL fills 49.75%, versus 89.64% for the behavior controller. Both learned proposals have considerably higher costs in this compact setup. Appending the robust screen to BC still reduces its within-bound violations from 15.6% to 0.0%; it does not repair the BC controller's low service. This is an important negative result: a valid safety filter cannot create a high-quality replenishment policy from a poor proposal.

The mean operating cost and service response of the screened behavior policy remain more informative than emissions alone. When the assumption set is increased, the filter admits fewer candidate trips late in the horizon. In the severe scenario, unfiltered behavior preserves high immediate service but inevitably overspends the cap in this evaluation, whereas robust variants trade part of that service for substantially lower violation risk. This mechanism is visible in Fig. 2. The observed cost increases are primarily downstream backlog consequences of limited replenishment, and should not be interpreted as direct taxes on carbon.

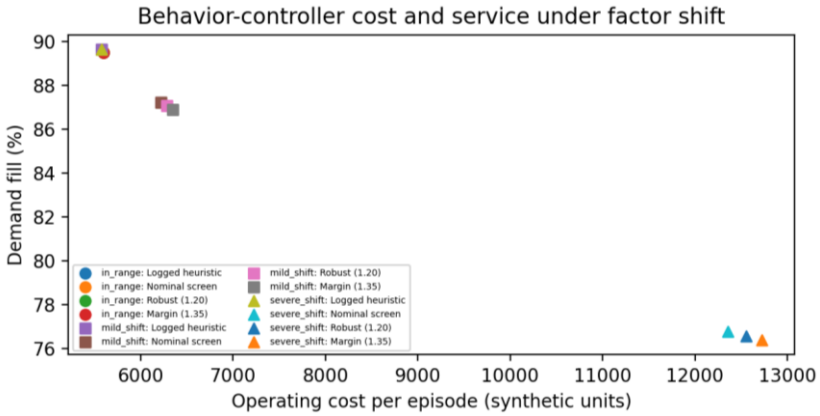


Figure 2. Costs and fill rates of four behavior-based controllers across within-bound, mild and severe factor shifts.

6.3. Carbon-cap sensitivity

Table IV. Zero-Retraining Cap Sensitivity (Within-Bound Errors)

Cap kg	Policy	Cost	Fill %	Viol. %
3900	Nominal screen	6,300	87.0	12.8
3900	Robust (1.20)	6,391	86.8	0.0
3900	BC nominal	29,280	50.3	7.2
3900	BC + robust	29,301	50.3	0.0
4700	Nominal screen	5,575	89.6	0.8
4700	Robust (1.20)	5,574	89.6	0.0
4700	BC nominal	29,024	50.1	0.0
4700	BC + robust	29,024	50.1	0.0

The same trained BC and support-CQL models are reused without additional training when the cap changes. The robust behavior policy has no observed cap violation at either 3900 or 4700 kg, as expected under the same within-bound multiplier assumption. At the tighter cap, its fill is 86.78%; at the relaxed cap it is 89.64%. Budget awareness in this experiment is a deployment filter, not a claim that the learned Q weights generalize to arbitrary new cost models.

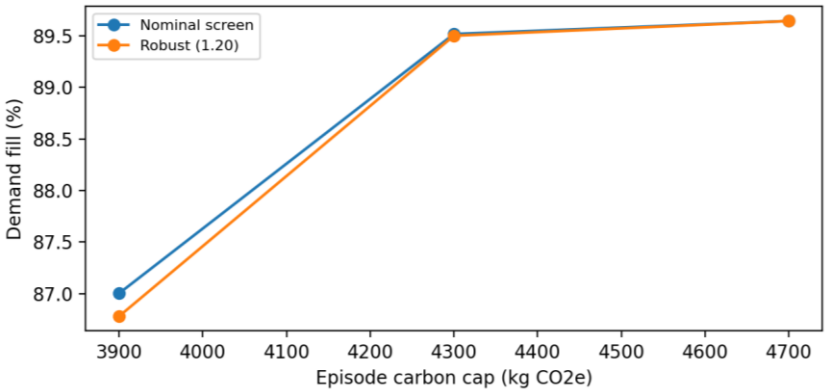


Figure 3. Zero-retraining service response to tighter and looser realized carbon budgets.

6.4. Paired Evaluation and Independent-Unit Inference

For paired behavior screening, the mean difference robust minus nominal in within-bound violation frequency is -5.2 percentage points. In mild shift it is -17.6 points; in severe shift it is -25.6 points. The seed-wise differences for each of these three contrasts share one sign, giving an exact two-sided sign-flip p=0.0625. This finite-sample test supports a consistently observed direction across five independent training runs without implying significance beyond what five independent datasets can resolve.

A stronger comparison is possible here than with a single coupled simulator RNG: each evaluation episode uses identical pre-generated day-by-day demands and multiplier errors for all competing actions. Thus, an observed difference between two screens does not result from a policy using a different number of draws and unintentionally changing the subsequent demand sequence. Paired evaluation still does not confer external validity because every demand distribution, cost coefficient, and emissions equation is synthetic.

6.5. Joint Safety–Service Success

A raw carbon violation rate is not sufficient to distinguish an environmentally disciplined policy from one that simply stops replenishing. We therefore computed a second, prespecified diagnostic on the already saved test trajectories: joint success requires realized total emissions at or below B and at least 85% immediate-demand fill. The fill threshold is a transparent synthetic evaluation criterion, rather than a purported industrial service-level agreement. No policy was retrained or selected using this threshold. Joint success is determined once per episode from the same cost, demand, and realized carbon records that generate Tables I–IV.

Within the bounded range, joint success increases from 64.4% for unfiltered behavior to 82.0% with nominal screening and 84.8% with a 1.20 envelope. The gain occurs because the filters eliminate cap violations without materially weakening the strong heuristic proposal. In mild shift, the 1.20 envelope attains 68.0% joint success, compared with 55.2% under the point screen; the more conservative 1.35 envelope is at 67.2%. These small differences show why margin selection should not be based on carbon violations alone. In severe shift, all screened heuristics have less than 10% joint success: under strong factor misspecification, further carbon restrictions eventually damage short-horizon demand fill. The BC-based method has less than 1% joint success in every condition, despite its relatively low emissions.

Table V. Joint Emission Compliance And 85% Fill Success

Factor	Policy	Joint %	Fill %
In-bound	Logged heuristic	64.4%	89.6%
In-bound	Nominal screen	82.0%	89.5%
In-bound	Robust (1.20)	84.8%	89.5%
In-bound	Margin (1.35)	84.8%	89.5%
In-bound	BC + robust	0.8%	50.3%
Mild shift	Logged heuristic	9.2%	89.6%
Mild shift	Nominal screen	55.2%	87.2%
Mild shift	Robust (1.20)	68.0%	87.1%
Mild shift	Margin (1.35)	67.2%	86.9%
Mild shift	BC + robust	0.8%	50.3%
Severe shift	Logged heuristic	0.0%	89.6%
Severe shift	Nominal screen	7.2%	76.8%
Severe shift	Robust (1.20)	8.0%	76.6%
Severe shift	Margin (1.35)	9.2%	76.4%
Severe shift	BC + robust	0.8%	47.9%

6.6. What the Fixed Log Actually Covers

Table VI. Training-Log Coverage By Independent Seed

Seed	Transitions	Actions seen	Cover %
0	4800	147/159	92.5
1	4800	146/159	91.8
2	4800	144/159	90.6
3	4800	144/159	90.6
4	4800	141/159	88.7

The five offline datasets each contain 4,800 transitions, yet they cover between 141 and 147 of the 159 possible action indices. This is coverage of the discrete index set only: even when an action is observed once, its value in a particular inventory state remains weakly identified. The behavior policy is concentrated near its order-up-to targets; actions involving combinations of upstream ordering, retailer allocations, and freight mode that are rarely selected by the logged mixture can therefore have poor Bellman targets. The support-CQL model regularizes action values, but the relatively short fit and narrow state-action coverage still yield a weak operational controller. The screen has no means of restoring this missing information; its narrower contribution is conditional carbon feasibility after a proposal is supplied.

We deliberately retain the failed learned policies in Tables I–III. Their roughly 50% fill rates mean that aggregate emissions and violation counts alone would give a misleading impression of good control. The behavior clone was trained as a 159-class classifier, not as three separate replenishment subdecisions. A factorized or autoregressive action

representation might use the fixed log more efficiently, particularly when high-dimensional combinations are sparse. That is a plausible next experiment rather than a tested contribution of this study. Reporting these failures also prevents a strong carbon-screening result from being mistaken for evidence that the full offline optimization problem has been solved.

7. Robustness, Identifiability, And Limitations

7.1. What the Bound Does and Does Not Certify

The deterministic CA-CQL screen ensures modeled-ledger compliance when its input emissions are exact. The uncertain extension certifies the realized ledger only under an upper-bounded error model. The proof does not require the value network to be calibrated and survives even when BC or Q learning performs poorly; it requires only accurate pre-dispatch upper estimates and nonnegative emissions. This modularity is a practical strength but not a guarantee of service. A zero-emission action can always keep the budget, so the high-fill behavior baseline is necessary to establish that a resulting policy remains operationally useful.

The 1.35 margin is a sensitivity control, not a distribution-free statistical confidence interval. Its success under [0.9, 1.3] follows from set containment, while its 6.4% observed severe-shift violation rate demonstrates why it cannot be sold as protection against [1.1, 1.5]. A deployment system would need factor estimates and uncertainty sets that track actual fleet, route, load, and scope differences. Historical activity logs may support upper-envelope estimation, but a principled calibration design for temporally shifted fleets would need separate evidence and validation.

7.2. Proposal Quality, Carbon Accounting, And Reproducibility

The numerical network is small, with one product and three stores, and the transport emissions are hypothetical action-level kg CO2e rather than measured driving data. The simulator does not optimize stop order, conduct vehicle-level route planning, account for unmodeled refrigeration, or include warehouse, production, or manufacturing emissions. The hard ceiling concerns only the modeled dispatch ledger. Our action projection is a discrete feasibility repair with a weighted distance objective; alternative repair weights or anticipated demand may improve cost, but no optimality guarantee is inferred from the screened result.

The learned-policy ablation exposes a separate weakness: 300 updates on each 4,800-transition synthetic dataset did not produce competitive overall inventory control, despite using a behavior-support term with conservative value learning. The proposal failure should not be generalized to CQL itself, whose full algorithm and hyperparameter selection are broader. A realistic end-to-end offline RL contribution would need richer logged actions, independent hyperparameter validation, larger and more heterogeneous inventories, and comparisons with stronger offline inventory baselines. Our supported claim is the effect of the implemented deployment filter within this benchmark.

All five compressed offline datasets, five trained checkpoints, episode-level outcomes, diagnostics, paired tests, training configuration, exact source code, and a seed-0 action-screen audit are retained. The audit records for each selected dispatch its nominal emission, factor-adjusted upper estimate, carbon balance, chosen index, and actual realized footprint.

7.3. Sequential Risk Under Bounded and Misspecified Factors

A deterministic screen differs sharply from an expected-carbon penalty and from a marginal prediction interval. Penalizing mean emissions in the policy objective can reduce average intensity while leaving a positive chance of a rare high-cost dispatch when little budget remains. Likewise, a marginal 95% upper confidence estimate for each dispatch would not, by itself, guarantee that all 40 dispatches satisfy the upper bound simultaneously. With independent errors, the probability of zero failures at all 40 periods would depend on the joint error law; without independence, a union bound requires one-step error probabilities chosen to control the desired familywise probability. Our selected 1.20 coefficient is instead stipulated to be a deterministic upper endpoint of the within-bound synthetic simulation, so there is no estimated coverage level to infer from the test set.

The severe-shift evaluation is a direct counterexample to extending that logic without validating the upper envelope. Here the true multiplier can reach 1.50 while the proposed robust constraint reserves only 1.20 times nominal emissions. Earlier reductions in transport activity may create some useful slack, and the recorded violation rate of 18.8% is considerably below the point-screen rate of 44.4%, but the guarantee itself is gone. Increasing the multiplier to 1.35 lowers the observed severe-shift rate to 6.4%, yet this factor is still below the simulated maximum and therefore retains no formal pathwise certificate. The numerical results and theoretical condition are consistent because they answer different questions: observed safety in finitely many scenarios and logical safety for every realization in a specified set.

7.4. Carbon Ledger Timing and Audit Reproducibility

The simulated remaining budget uses realized C_t when the next action is chosen. This assumes that prior freight emissions have already been reconciled by the decision time. An operational company may instead receive carrier fuel reports, empty-mileage adjustments, or subcontractor data after a delay. In that case, the correct remaining committed budget cannot be computed by merely summing emissions already reported; the controller must also reserve an upper exposure for dispatched but unresolved movements. The same feasibility induction can be recovered if these outstanding exposures are bounded and deducted, but that variant is not evaluated here.

A dispatch-level screen replay checks 180,000 selected steps across 4,500 policy episodes, including equality of the saved final ledger and recomputed episode outcomes. A specific nominal-screen failure is identified in seed 0, evaluation episode 13, day 39. At selection time 121.62 kg CO2e remain. The screen permits action index 58 with nominal estimated emissions of 121.00 kg; its realized emission is 129.20 kg and the ledger consequently finishes 7.58 kg above the budget. The robust factor 1.20 would require 145.20 kg of headroom for this same action and reject it in this state. The counterexample is reproduced in the dispatch-level CSV, not constructed by adjusting a reported result.

A useful implementation audit records at least the proposed action, selected action, physical mask, assumed upper

multiplier, nominal action footprint, remaining verified budget, and realized footprint. The released seed-0 trace contains action indices, upper-screen test values, factor realizations, day-wise demand, and cumulative realized emissions for 40 periods of the nominal, 1.20, and 1.35 behavioral screening rules. It permits a reviewer to check the exact inequalities without relying on plotted averages. The five full policy checkpoints and five compressed logs are also preserved. All numerical settings, including unit-valued synthetic costs and trip coefficients, are in the released source code, making the benchmark a clearly limited but inspectable mechanism study.

8. Conclusion

An exact remaining-budget carbon screen can prevent a controller from overspending its predicted ledger yet still permit real-world carbon overruns when dispatch emissions are uncertain. We extended the pathwise screen of CA-CQL with a bounded factor and a behavior-preserving feasible-action projection. Five seeded offline datasets and 250 test episodes per uncertainty regime show a clean within-bound elimination of violations with minimal mean service loss; the same controller incurs nonzero violations once the factor uncertainty moves beyond its declared upper bound. Logged behavior, behavior cloning, and conservative Q-learning comparisons show that feasibility and supply-chain service are distinct requirements. The approach supports an auditable robustness interface, while larger action spaces, actual freight measurements, and calibrated time-varying emissions bounds remain important future research directions.

References

[1] Clark, A. J., & Scarf, H. (1960). Optimal policies for a multi-echelon inventory problem. Management Science, 6(4), 475–490. https://doi.org/10.1287/mnsc.6.4.475

[2] Campbell, A. M., Clarke, L., Kleywegt, A. J., & Savelsbergh, M. W. P. (1998). The inventory routing problem. In Fleet Management and Logistics (pp. 95–112). Springer. https://doi.org/10.1007/978-1-4615-5755-5_4

[3] Li, P., Bian, M., & Lin, Y. (2026). Carbon-aware offline reinforcement learning for joint inventory replenishment and transportation scheduling in multi-echelon supply chains. Highlights in Business, Economics and Management, 69, 251–262. https://doi.org/10.54097/p40ecd94

[4] Kumar, A., Zhou, A., Tucker, G., & Levine, S. (2020). Conservative Q-learning for offline reinforcement learning. Advances in Neural Information Processing Systems, 33, 1179–1191.

[5] Fujimoto, S., Meger, D., & Precup, D. (2019). Off-policy deep reinforcement learning without exploration. Proceedings of the International Conference on Machine Learning, 2052–2062.

[6] Kumar, A., Fu, J., Soh, M., Tucker, G., & Levine, S. (2019). Stabilizing off-policy Q-learning via bootstrapping error reduction. Advances in Neural Information Processing Systems, 32.

[7] Fu, J., Kumar, A., Nachum, O., Tucker, G., & Levine, S. (2020). D4RL: Datasets for deep data-driven reinforcement learning. arXiv preprint arXiv:2004.07219.

[8] Altman, E. (1999). Constrained Markov Decision Processes. CRC Press.

[9] Achiam, J., Held, D., Tamar, A., & Abbeel, P. (2017). Constrained policy optimization. Proceedings of the International Conference on Machine Learning, 22–31.

[10] Bektaş, T., & Laporte, G. (2011). The pollution-routing problem. Transportation Research Part B, 45(8), 1232–1250. https://doi.org/10.1016/j.trb.2011.02.004

[11] Demir, E., Bektaş, T., & Laporte, G. (2014). A review of recent research on green road freight transportation. European Journal of Operational Research, 237(3), 775–793. https://doi.org/10.1016/j.ejor.2013.12.033

[12] Cheng, C., Qi, M., Wang, X., & Zhang, Y. (2016). Multi-period inventory routing problem under carbon emission regulations. International Journal of Production Economics, 182, 263–275. https://doi.org/10.1016/j.ijpe.2016.09.001

[13] Kleywegt, A. J., Nori, V. S., & Savelsbergh, M. W. P. (2002). The stochastic inventory routing problem with direct deliveries. Transportation Science, 36(1), 94–118. https://doi.org/10.1287/trsc.36.1.94.574

[14] Kleywegt, A. J., Nori, V. S., & Savelsbergh, M. W. P. (2004). Dynamic programming approximations for a stochastic inventory routing problem. Transportation Science, 38(1), 42–70. https://doi.org/10.1287/trsc.1030.0041

[15] Campbell, A. M., & Savelsbergh, M. W. P. (2004). A decomposition approach for the inventory-routing problem. Transportation Science, 38(4), 488–502. https://doi.org/10.1287/trsc.1030.0054

[16] Trudeau, P., & Dror, M. (1992). Stochastic inventory routing: Route design with stockouts and route failures. Transportation Science, 26(3), 171–184. https://doi.org/10.1287/trsc.26.3.171

[17] Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction (2nd ed.). MIT Press.

[18] World Resources Institute & World Business Council for Sustainable Development. (2011). Corporate Value Chain (Scope 3) Accounting and Reporting Standard.

[19] Smart Freight Centre. (2024). Global Logistics Emissions Council Framework for Logistics Emissions Accounting and Reporting (Version 3.1).

[20] UK Department for Energy Security and Net Zero & Department for Environment, Food and Rural Affairs. (2026). Greenhouse Gas Reporting: Conversion Factors 2026.

[21] Gibbs, I., & Candès, E. (2021). Adaptive conformal inference under distribution shift. Advances in Neural Information Processing Systems, 34.

[22] Dong, Q., & Yuan, Y. (2025). Data-driven distributionally robust supplier selection and order allocation problems considering carbon emissions. International Transactions in Operational Research, 32(2), 1119–1145. https://doi.org/10.1111/itor.13328