



Delft University of Technology

Document Version

Final published version

Licence

CC BY-NC-ND

Citation (APA)

Saifullah, M., Papakonstantinou, K. G., Bhattacharya, A., Stoffels, S. M., & Andriotis, C. P. (2026). Multi-agent deep reinforcement learning with centralized training and decentralized execution for transportation infrastructure management. *Computer-Aided Civil and Infrastructure Engineering*, 51, Article 100175. <https://doi.org/10.1016/j.cacaie.2026.100175>

Important note

To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright

In case the licence states "Dutch Copyright Act (Article 25fa)", this publication was made available Green Open Access via the TU Delft Institutional Repository pursuant to Dutch Copyright Act (Article 25fa, the Taverne amendment). This provision does not affect copyright ownership.
Unless copyright is transferred by contract or statute, it remains with the copyright holder.

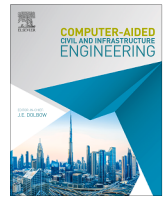
Sharing and reuse

Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy

Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.



Multi-agent deep reinforcement learning with centralized training and decentralized execution for transportation infrastructure management

M. Saifullah^a, K.G. Papakonstantinou^{a,*}, A. Bhattacharya^a, S.M. Stoffels^a, C.P. Andriotis^b

^a Department of Civil and Environmental Engineering, The Pennsylvania State University, University Park, PA 16802, USA

^b Faculty of Architecture & the Built Environment, Delft University of Technology, 2628 BL, Delft, the Netherlands

ARTICLE INFO

Keywords:

Cross-asset infrastructure management
Transportation networks
Stochastic deterioration
Partially observable Markov decision processes
Decentralized multi-agent planning
Constrained deep reinforcement learning

ABSTRACT

Life-cycle management of large-scale transportation systems is a computationally intensive task. It requires determining a sequence of inspection and maintenance decisions to minimize long-term risks and costs while dealing with multiple uncertainties and constraints that lie in high-dimensional spaces. Traditional approaches, such as static age- or condition-based maintenance and risk-based or periodic inspection plans, have been widely applied but often suffer from limitations related to optimality, scalability, and the ability to properly handle uncertainty in the loop of optimization. Moreover, many existing methods rely on unconstrained formulations that overlook critical hard and soft operational constraints. We address these issues in this work by casting the optimization problem within the framework of constrained Partially Observable Markov Decision Processes (POMDPs), which provide a robust mathematical foundation for stochastic sequential decision-making under observation uncertainties, in the presence of risk and resource limitations. To tackle the high dimensionality of state and action spaces, we propose DDMAC-CTDE, a Deep Decentralized Multi-Agent Actor-Critic (DDMAC) reinforcement learning architecture with Centralized Training and Decentralized Execution (CTDE). To demonstrate the utility of the proposed framework, we also develop a new comprehensive benchmark environment representing an existing transportation network in Virginia, U.S., with heterogeneous pavement and bridge assets undergoing nonstationary degradation. This environment incorporates multiple practical constraints related to budget limits, performance guidelines from State and Federal agencies, traffic delays, and risk considerations. On this benchmark, DDMAC-CTDE consistently outperforms standard transportation management baselines, producing more reliable, cost- and risk-efficient policies that respect operational constraints. Together, the proposed framework and benchmark provide (i) a scalable, constraint-aware methodology for sequential decision-making under uncertainty, and (ii) a realistic, rigorous testbed for comprehensive evaluation of Deep Reinforcement Learning (DRL) solutions for transportation infrastructure management.

1. Introduction

The effective management of transportation infrastructure remains a significant challenge, requiring the development of advanced inspection and maintenance policies that address stochastic degradation processes while adhering to stringent performance and resource constraints. This issue is particularly pressing in light of infrastructure conditions in many countries worldwide, as also revealed by the 2025 ASCE Infrastructure Report Card (American Society of Civil Engineers, 2025). Several economic analyses also highlight a substantial funding deficit in preserving infrastructure to acceptable societal standards, both at the Federal and State levels (American Society of Civil Engineers (2025); VDOT (2022)). For example, and only illustratively, the Virginia Department

of Transportation (VDOT) reported that 50% of the state's bridges have exceeded their intended service lives and the projected costs to replace these aging structures over the next 50 years are estimated to be five times greater than the anticipated available funding (VDOT (2022)). These facts further underscore the urgent need for innovative and resource-efficient strategies to ensure the sustainability, safety, and performance of transportation infrastructure systems.

Intelligent inspection and maintenance (I&M) planning is key to the efficient allocation of economic and other resources in such systems. Developing optimal policies in complex and uncertain environments poses significant computational challenges, stemming from the heterogeneity of asset classes, the large population of components leading to

* Corresponding author.

Email address: kpapakon@psu.edu (K.G. Papakonstantinou).

<https://doi.org/10.1016/j.cacaie.2026.100175>

Received 18 March 2026; Received in revised form 4 August 2026; Accepted 5 August 2026

Available online 11 August 2026

1093-9687/© 2026 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

vast state and action spaces, the imperfect and unreliable observations, the scarcity of available resources, and the multiple associated risks. Despite progress, there remains a shortage of scalable computational frameworks that can jointly handle features such as (i) online and offline data-driven learning, (ii) imperfect information support, (iii) stochastic action outcomes, and (iv) optimization of long-term goals under multiple constraints (e.g., safety targets or resource constraints). Computational challenges, such as the curse of dimensionality and history, further complicate this objective Bellman (1957); Pineau et al. (2003). Overcoming all these challenges necessitates the development of advanced computational frameworks capable of integrating these diverse elements while ensuring scalability and computational feasibility. At the same time, standardized, sufficiently realistic, detailed, and rigorous benchmarks that can reflect the dynamic, stochastic, and constrained nature of transportation infrastructure management are also currently lacking in the literature. This impedes fair comparison, reproducibility, and collective progress. Developing comprehensive planning environments, with realistic heterogeneity, observation noise, resource limits, and risk targets, would enable systematic evaluation and accelerate the design of appropriate, safe, and scalable I&M policies.

Traditional computational methods face significant limitations in addressing these multifaceted challenges within I&M optimization. Over the years, numerous advanced techniques have been proposed, including genetic algorithms, threshold-based heuristics, risk-informed approaches, renewal processes, and decision-tree frameworks Bocchini and Frangopol (2011); Chang et al. (2023); Mendoza et al. (2021); Nguyen et al. (2015); Sánchez-Silva et al. (2016); Yang and Frangopol (2019). Among these, risk-based approaches have garnered particular attention, yielding practical solutions Bismut and Straub (2021); Han and Frangopol (2022, 2023); Yang and Frangopol (2021), but they often struggle with large-scale systems, computational inefficiencies, and sub-optimal scalability. Additionally, many rely on static formulations that fail to capture the dynamic and stochastic characteristics of infrastructure deterioration and maintenance needs, limiting their applicability for real-world, non-stationary systems. These shortcomings are also evident in transportation asset management studies, such as Azad et al. (2016); Chu and Chen (2012); Frangopol and Liu (2007a); Saydam and Frangopol (2015). A broader review of multi-asset approaches, including several transportation-focused studies, is presented in Petchrompo and Parlikad (2019). Although such multi-asset frameworks provide a more holistic perspective on network-level optimization, as emphasized in Petchrompo and Parlikad (2019) and the extensive references therein, current state-of-the-art methods remain unable to inclusively address all the previously mentioned challenges and the core concepts (i)–(iv) above. In addition, none of the currently available studies offer a truly comprehensive planning environment capable of managing a large number of components with dynamic, adaptive planning needs, diverse optimization objectives, and various uncertainties, all while respecting real-world operational performance and resource constraints.

Stochastic optimal control approaches based on Markov Decision Processes (MDPs) and Partially Observable Markov Decision Processes (POMDPs) can tackle many of the aforementioned challenges Frangopol and Liu (2007b); Papakonstantinou and Shinozuka (2014a); Papakonstantinou et al. (2018). POMDPs' dynamic programming principles alleviate the curse of history and enable adaptive reasoning in the presence of noisy real-time data Kaelbling et al. (1998). Numerous research efforts, including Memarzadeh and Pozzi (2016); Memarzadeh et al. (2015); Papakonstantinou and Shinozuka (2014b, 2014c); Papakonstantinou et al. (2016); Schöbi and Chatzi (2016), have explored and demonstrated the successful use of POMDPs and their efficient performance in I&M planning problems. In Morato et al. (2022), POMDPs are also combined with dynamic Bayesian networks, to balance the costs of maintenance actions and observations against the risk of structural failures, whereas in (Arcieri et al., 2023) a robust POMDP approach is suggested, considering model uncertainty. POMDPs have also been used for pavement management, as in (Guillaumot et al.,

2003; Madanat & Ben-Akiva, 1994; Madanat, 1993), while in Zhang et al. (2023) a stochastic condition-based I&M optimization is suggested, and in Shon and Lee (2021) a bottom-up dynamic programming solution is presented for the optimization problem. There are many ways to efficiently solve POMDPs to global optimality, e.g., Lovejoy (1991); Papakonstantinou et al. (2018); Pineau et al. (2003); Shani et al. (2013); Spaan and Vlassis (2005). However, unfortunately, most traditional algorithms still fail to efficiently scale to large systems with multiple components.

To address relevant scalability issues, neural network parameterizations can be utilized, through the field of Deep Reinforcement Learning (DRL) (Andriotis & Papakonstantinou, 2019a; Mnih et al., 2015; Schulman et al., 2017). Several works have thus been recently published, examining DRL performance and capabilities for I&M planning problems. For example, in Lei et al. (2022), DQN learning is used for life-cycle maintenance planning of deteriorating bridges under different budget constraints, while in Du and Ghavidel (2022) DQN is used for sequential maintenance decisions for highway bridge portfolios. In Hamida and Goulet (2023) hierarchy in actions and a Dueling DQN are used for maintenance planning of bridges, whereas a Double DQN with a prioritized replay buffer is used in Mohammadi and He (2022) for rail renewal and maintenance planning, and a POMDP-DRL framework for inference and robust solutions is presented in Arcieri et al. (2024), together with an application to railway optimal maintenance. Several other notable relevant approaches can be seen in (Lai et al., 2024; Lei et al., 2023; Sresakoolchai & Kaewunruen, 2023; Yang, 2022a; Yao et al., 2020). Other diverse applications include aircraft maintenance (Lee & Mitici, 2023), autonomous vehicles Zhang et al. (2023), structural monitoring Pan et al. (2024), and many more.

Most DRL approaches for transportation I&M planning are designed within single-agent frameworks, which face significant scalability challenges in multi-component systems due to the exponential growth of joint state and action spaces Fard and Yuan (2025). Addressing this issue, one of the early Multi-Agent DRL (MADRL) techniques for I&M, the Deep Centralized Multi-agent Actor-Critic (DCMAC) algorithm, has been proposed in Andriotis and Papakonstantinou (2019a, 2019b), which is a subset of the larger family of actor-critic DRL approaches Degris et al. (2012); Wang et al. (2016). DCMAC addresses partial observability by making use of belief-state MDPs, and therefore operates directly on the posterior probabilities of system states given previous actions and observations. DCMAC has also been applied to address a bilevel optimization problem of scheduling roadway improvements while considering traffic impacts associated with maintenance in (Zhou et al., 2022). (Yang, 2022b) suggested a DCMAC variant by incorporating an entropy term, applying the approach to the life-cycle maintenance of bridge girder systems, while (Zhang et al., 2024) also utilized a centralized multi-agent framework using PPO (Schulman et al., 2017) to optimize life-cycle maintenance decisions for bridge networks, considering the interdependencies between bridge components.

To enable decentralized parameter training, the Deep Decentralized Multi-agent Actor-Critic (DDMAC) is proposed in Andriotis and Papakonstantinou (2021), where a decentralized independent actor unit now represents each component, and their outputs are utilized to build a centralized value function, which updates the actor and critic networks using policy gradients and temporal difference error estimates. DDMAC is also used to compute optimal system level I&M policies, for systems with correlated fatigue deterioration processes among components Morato et al. (2023), and a Maintainer-Inspector architecture for sequential actions is presented in Andriotis and Papakonstantinou (2022); Saifullah et al. (2023), based on the DDMAC framework, found to be notably efficient for settings with large combinations of available I&M actions. Beyond DDMAC, alternative multi-agent deep reinforcement learning algorithms have also led to advancements in maintenance policies and cost-effectiveness. In Do et al. (2024); Nguyen et al. (2022b, 2022a); Yao et al. (2024) MADRL methods based on the value decomposition framework of QMIX Rashid et al. (2020) are

presented, while Lee and Song (2023) presents a parallelized multi-agent deep Q-network algorithm for risk-informed lifeline system maintenance. The Value-Decomposition Actor-Critic (VDAC) algorithm Su et al. (2021) is employed in Su et al. (2022), for preventive maintenance policies of manufacturing systems, and Wang et al. (2016) devises multi-agent maintenance strategies for resource-constrained flow line systems. Similarly, Rodríguez et al. (2022) capitalizes on MADRL for predictive maintenance on parallel machines, showcasing substantial performance enhancements. Several MADRL approaches and benchmarks for asset management can also be seen in Leroy et al. (2023).

In this work, we advance research on deep reinforcement learning for transportation infrastructure applications by developing the first multi-asset transportation infrastructure management environment accounting for realistic, high-fidelity system modeling and several, diverse, deterministic and probabilistic agency constraints. This environment is designed for optimizing I&M under various operational metrics and uncertainties. Next, to enable and scale the DDMAC architecture for even larger systems, we introduce it here to the Centralized Training and Decentralized Execution (CTDE) paradigm, a notable framework in cooperative multi-agent reinforcement learning (MARL) built on the Decentralized-POMDP (Dec-POMDP) formalism Bernstein et al. (2002); Oliehoek and Amato (2016). This DDMAC-CTDE architecture provides now only local information as input to decentralized actors Bhattacharya et al. (2025), reflecting the condition states of individual components. This new design substantially enhances scalability capabilities while preserving centralized gradient information flow through a single critic, consistent with the original DDMAC framework. The proposed implementation is evaluated here in a realistic, large-scale setting. We formulate the POMDP optimization problem within the benchmark environment representing an existing real-world transportation network at Hampton Roads, Virginia, U.S.A, with two classes (pavements and bridges) and multiple subclasses of assets under numerous resource, performance, and risk constraints. A detailed modeling environment of both pavements and bridges is developed based on their corresponding condition state indices. Stationary and non-stationary Markovian transition probabilities are obtained for different classes and subclasses of assets. Four types of maintenance actions are considered, namely Do-Nothing, Minor Repair, Major Repair, and Reconstruction. Observation actions are, in turn, defined as no-inspection, low-fidelity, and high-fidelity inspection for both asset types. The maintenance and inspection action costs and effects are obtained from actual data and complemented by up-to-date literature, where needed. Extensive numerical results are presented demonstrating the effectiveness and applicability of the proposed DDMAC-CTDE formulation for this type of large-scale, extremely challenging system optimization problems, with derived solutions outperforming state-of-the-art heuristics and currently in-place policies by a significant margin, while satisfying the imposed performance and budget constraints. Overall, this work importantly delivers (i) a scalable computational framework for DRL-driven I&M, and (ii) a realistically complex benchmark for fair, demanding, reproducible evaluation. These two contributions together lay a solid and necessary foundation for even further future advances in real-world DRL-driven management and engineering decision-making, broadly defined.

2. Background

2.1. Partially observable Markov decision processes

MDP is represented by a tuple $\langle S, A, \mathbb{P}, C, \gamma \rangle$ for environment E , where S is the state space, A is a control action space, $\mathbb{P}$ is the environment transition model, C denotes the cost (or negative reward) model, and $\gamma \in [0, 1]$ is a scalar discount factor. At each decision step t , given the current state $s_t \in S$, the agent takes an action $a_t \in A$, incurs a cost as a result of this state and action, $c(s_t, a_t)$, and the system transitions to state $s_{t+1} \in S$ based on the Markovian probability $p(s_{t+1} | s_t, a_t)$. The agent chooses actions according to its policy π , which can be deterministic, $\pi(s_t): S \rightarrow A$ or stochastic, i.e., a distribution over actions conditioned

on s_t , i.e., $\pi(a_t | s_t): S \rightarrow P(A)$. Policy $\pi(s_t)$ is thus a scalar value for a deterministic policy, while for a stochastic policy, π is a probability mass and density function for discrete and continuous action spaces, respectively. The corresponding cost function $C_t = \sum_{i=t}^T \gamma^{i-t} c(s_i, a_i)$ is the total cost incurred from time step t to the end of the planning horizon T under the policy π . Further, we can define the action-value function, Q^π , as the expected cost over all potential future states and actions, conditioned on the current state-action pair, s_t, a_t as:

$$\begin{aligned} Q^\pi(s_t, a_t) &= \mathbb{E}_{s_{t+1} \sim E, a_{t+1} \sim \pi} [C_t | s_t, a_t] \\ &= c(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim E, a_{t+1} \sim \pi} [Q^\pi(s_{t+1}, a_{t+1})] \end{aligned} \quad (1)$$

Similarly, a value function of a policy π is the cost expectation from state s_t . It can also be defined as the expectation of the action-value function over all possible actions at the current step:

$$V^\pi(s_t) = \mathbb{E}_{a_t \sim \pi} [Q^\pi(s_t, a_t)] \quad (2)$$

The goal of the agent is to select the optimal sequence of actions that minimize the value function, constituting the optimal policy π^* . For the optimal policy, with a given model of transitions $p(s_{t+1} | s_t, a_t)$ and the aid of Eq. (1), the optimal action-value $Q^{\pi^*}(s_t, a_t)$ follows the Bellman equation Bellman (1957), and can be obtained as:

$$Q^{\pi^*}(s_t, a_t) = c(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim E} \left[\min_{a \in A} Q^{\pi^*}(s_{t+1}, a_{t+1}) \right] \quad (3)$$

Similarly, the optimal value function $V^{\pi^*}(s_t)$ can be estimated as:

$$\begin{aligned} V^{\pi^*}(s_t) &= \min_{a \in A} \{ Q(s_t, a_t) \} \\ &= \min_{a \in A} \left\{ c(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim E} \left[\min_{a \in A} Q^{\pi^*}(s_{t+1}, a_{t+1}) \right] \right\} \\ &= \min_{a \in A} \left\{ c(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1} \sim E} [V^{\pi^*}(s_{t+1})] \right\} \\ &= \min_{a \in A} \left\{ c(s_t, a_t) + \gamma \sum_{s_{t+1} \in S} p(s_{t+1} | s_t, a_t) V^{\pi^*}(s_{t+1}) \right\} \end{aligned} \quad (4)$$

The objective function of MDP can thus be described by Eq. (3) or (4), which can be solved using value iteration, policy iteration, or linear programming formulations, as presented in Papakonstantinou and Shinozuka (2014b).

A natural extension to an MDP is the partially observable MDP (POMDP), which is a generalization of the case when the environment is not perfectly observable. This is relevant to a wide variety of engineering applications where the available information is incapable of revealing the true state of the system. Analogous to MDPs, in the POMDP framework the decision-maker/agent starts at a hidden state, s_t , at a time step, t , takes an action a_t , incurs a cost, c , transitions to the next hidden state, s_{t+1} , and receives an observation, $o_{t+1} \in \Omega$, based on the observation probability model, $p(o_{t+1} | s_{t+1}, a_t)$ for a given state and action. More formally, a POMDP is a tuple $\langle S, A, \mathbb{P}, \Omega, \mathbb{O}, C, \gamma \rangle$ where S , A and Ω are sets of states, actions, and possible observations, respectively, and $\mathbb{P}$, $\mathbb{O}$, and C are the transition, observation, and cost models, respectively.

Due to the uncertainty associated with observations, the agent does not know the actual state, hence it can only form a belief $\mathbf{b}_t$ about its state, where $\mathbf{b}_t$ is a probability distribution over the set S of all possible discrete states. A Bayesian update can be used to calculate the new belief $\mathbf{b}_{t+1}$ for a given action and observation, Papakonstantinou and Shinozuka (2014b):

$$\begin{aligned} b(s_{t+1}) &= p(s_{t+1} | o_{t+1}, a_t, \mathbf{b}_t) \\ &= \frac{p(o_{t+1} | s_{t+1}, a_t)}{p(o_{t+1} | \mathbf{b}_t, a_t)} \sum_{s_t \in S} p(s_{t+1} | s_t, a_t) b(s_t) \end{aligned} \quad (5)$$

where probabilities $b(s_t)$, for all $s_t \in S$, form the belief vector $\mathbf{b}_t$ of length $|S|$, and the denominator of Eq. (5), $p(o_{t+1} | \mathbf{b}_t, a_t)$, is the standard

normalizing constant. As such, beliefs can be seen as alternative states of this environment, and POMDPs can be accordingly regarded as belief-MDPs. Further, the optimal value function V^{π^*} , following the optimal π^* , takes an expression similar to Eq. (3):

$$V^{\pi^*}(\mathbf{b}_t) = \min_{a_t \in A} \left\{ \sum_{s_t \in S} b(s_t) c(s_t, a_t) + \gamma \sum_{o_{t+1} \in \Omega} p(o_{t+1} | \mathbf{b}_t, a_t) V^{\pi^*}(\mathbf{b}_{t+1}) \right\} \quad (6)$$

Despite the similarity with MDPs, POMDPs are more difficult to solve. The new belief-state space is continuous rather than discrete, producing a $(|S| - 1)$ -dimensional simplex. The optimum value function has been proved however to be piece-wise linear and convex, and as such it can be represented by a finite number of affine hyperplanes (Sondik, 1971). Thus, the decision problem is reduced to choosing a finite number of vectors, commonly known as the α -vectors. To solve this problem, point-based algorithms seek to iteratively determine a set of α -vectors that can adequately support the belief space Kurniawati et al. (2008); Spaan and Vlassis (2005). However, such point-based solvers can become impractical for large POMDP problems, particularly for infrastructure systems with many components (Papakonstantinou et al., 2018). As already mentioned, the curse of dimensionality associated with POMDPs can be alleviated using a deep reinforcement learning framework.

2.2. Deep reinforcement learning

Reinforcement learning (RL) is a computational framework for evaluating and automating goal-directed learning and decision-making, and is especially well-suited for solving MDP/POMDP problems as it is mathematically structured around them Sutton and Barto (2018). RL is theoretically able to alleviate the curse of dimensionality related to the state space, either under model-free approaches that do not utilize prior environment information on transition dynamics, or model-based approaches based on an underlying transition model of the environment or hybrid approaches thereof.

Deep Reinforcement Learning (DRL) utilizes deep neural network representations to approximate the value functions or policies, demonstrating capabilities for discovering meaningful parameterizations in immense state spaces, outperforming human experts Sutton and Barto (2018); Sutton et al. (2000). DRL algorithms are generally categorized into value-based and policy-based learning methods. Value-based methods learn the state or state-action value function and determine the optimal action for each state Mnih et al. (2015); Sutton and Barto (2018); Van Hasselt et al. (2016). In contrast, policy-based methods directly learn a policy, $\pi : S \rightarrow A$ or $\pi : S \rightarrow P(A)$, using policy gradient techniques Williams (1992). Actor-critic methods combine both approaches by simultaneously learning a policy (actor) and a value function (critic), enhancing stability and convergence (Konda & Tsitsiklis, 2000). This hybrid strategy is widely adopted in DRL algorithms, including Proximal Policy Optimization (PPO) (Schulman et al., 2017), Asynchronous Advantage Actor-Critic (A3C) (Mnih et al., 2016), and Deep Deterministic Policy Gradient (DDPG) (Lillicrap et al., 2016).

Multi-Agent Deep Reinforcement Learning (MADRL) for optimal decision-making in managing infrastructure systems has also been introduced in the literature Andriotis and Papakonstantinou (2019a, 2021). However, MADRL methods have primarily been evaluated so far in environments with tens of agents or components. In this work, we aim to advance the application of MADRL by implementing it in a more complex, real-world scenario with nearly one hundred of heterogeneous components and several constraints. In the following section, we are detailing the MADRL formulation employed in our study.

3. DDMAC-CTDE

Designing optimal I&M policies for large-scale, long-horizon, multi-component systems is inherently challenging due to the exponential growth of joint state and action spaces with the number of components.

This combinatorial complexity renders traditional single-agent planning and learning algorithms computationally infeasible for such high-dimensional problems. Cooperative MADRL alleviates this complexity by addressing decision-making as a shared task among multiple autonomous agents. The agents optimize a shared long-term objective conditionally and independently in the same environment, thereby, making the joint policy scale linearly with the number of components. This paradigm is particularly well-suited for cross-asset life-cycle management of infrastructure systems, where coordinated planning of I&M actions across diverse components, such as bridges, pavements, and other assets is essential. By fostering cooperation among agents, this approach can potentially minimize overall risks and achieve various cost-related objectives, ensuring sustained performance and reliability of critical infrastructure.

In Andriotis and Papakonstantinou (2019a), the Deep Centralized Multi-agent Actor-Critic (DCMAC) approach was introduced to address multi-component decision-making challenges within the framework of off-policy actor-critic DRL. This approach leverages a replay buffer, a memory that stores past experiences collected by agents as they interact with their environment Degris et al. (2012); Sutton (1988). These stored experiences are utilized for stable batch training of the neural networks through random sampling at each training step Mnih et al. (2015). DCMAC uses policy gradient methods Sutton et al. (2000) to update the policy parameters and has shown significant life-cycle cost improvements compared to heuristically optimized policies Andriotis and Papakonstantinou (2019b). DCMAC is designed as a centralized agent with respect to access to the full state space across all agents, predicting decentralized actions for every agent in the system. While information centralization in principle benefits coordinated behavior, it can, in certain cases, hinder training efficiency. Parameterization over the full state space introduces significant computational complexity, particularly as the number of agents increases.

To alleviate some of the computational challenges, the DDMAC architecture, proposed in Andriotis and Papakonstantinou (2021), represents each agent with a separate actor/policy network. This design allows for a sparser policy parameterization by eliminating inter-agent interactions in the hidden layers, thereby improving training efficiency in certain cases. In Andriotis and Papakonstantinou (2021), DDMAC is further integrated into an algorithmic framework that supports planning under operational constraints, such as budget limitations. While DDMAC can offer improvements over DCMAC, it assumes that each agent has access to the entire system's state distribution. As a result, each actor receives the full state space as input, which may add a computational burden that is disproportionate to the expected performance gains. This is because each actor must optimize over the full state space, which again becomes increasingly computationally expensive as the number of agents grows. Furthermore, it also limits applicability in true decentralized execution settings, where agents need to make decisions based solely on local observations rather than the global state.

To be able to address the large systems examined in this work, we herein introduce a new variant of DDMAC, equipped with the Centralized Training and Decentralized Execution (CTDE) paradigm of cooperative reinforcement learning Foerster et al. (2018); Gronauer and Diepold (2022); Oliehoek et al. (2008); Rashid et al. (2018); Sunehag et al. (2017). This modified architecture, referred to as DDMAC-CTDE, builds upon the Decentralized-POMDP (Dec-POMDP) framework Oliehoek et al. (2008), modeling agents under fully decentralized policies, each conditioned, generally, on its corresponding component's access to local information/belief state. This ensures that agents operate independently during execution, eliminating the need for centralized control and enabling scalability to larger systems with a greater number of agents. DDMAC-CTDE is thus an adaptation of the original DDMAC architecture to a CTDE-compatible information structure, rather than a completely distinct MARL algorithm. In the original DDMAC, each

component is represented by a decentralized actor network, but each actor may still be conditioned on system-wide belief/state information, which can prove difficult to train and deploy in large-scale infrastructure systems. In the proposed DDMAC-CTDE formulation instead, each actor is now conditioned only on its local component belief and any relevant shared operational variables, such as budget and constraint states. This approach follows the standard CTDE principle, where centralized information guides training and system-level value estimation, while execution remains decentralized. CTDE is also an information-sharing paradigm, rather than a single algorithmic form, and can be implemented differently across MARL families, for example through value mixing in value-based methods or centralized critics in actor-critic methods. Hence, we here adapt the DDMAC actor-critic structure to CTDE while preserving the belief-state representation required by partial observability and the budget and performance constraints handling needed for infrastructure management, demonstrating it on a realistic, high-dimensional, multi-asset transportation benchmark. Related infrastructure MARL work has also presented relevant centralized and decentralized information structures for large bridge networks [Bhattacharya et al. \(2025\)](#).

The training process in the proposed framework employs a centralized critic, which has access to the full system belief state, guiding each agent's policy during training. This centralized critic evaluates the global state and joint behavior of all agents, ensuring that policy updates are informed through a global environment signal emitted through joint interaction among agents. However, the centralized critic is used exclusively during the training phase and is discarded during the execution or deployment phase, allowing for fully decentralized decisions at inference time. The centralized critic also helps mitigate the non-stationarity issue that arises in multi-agent settings, where simultaneous updates to agents' policies can make the environment's dynamics appear unstable, by leveraging global state information to provide a stable training signal, thereby reducing inconsistencies and accelerating convergence [Foerster et al. \(2018\)](#); [Lowe et al. \(2017\)](#).

This combination of Centralized Training and Decentralized Execution strikes a delicate balance between leveraging centralized information for effective training sparsifying the actor input and maintaining decentralized control for scalability. The CTDE framework is generally particularly powerful in cooperative multi-agent reinforcement learning, where agents share similar tasks and require coordination. By enabling agents to optimize actions within their local observation space while benefiting from centralized guidance during training, CTDE ensures scalability and robustness in environments with a large number of agents and complex interdependencies [Gronauer and Diepold \(2022\)](#). The suitability of the CTDE approach, however, can also depend on the coupling structure of the examined systems. In the benchmark problem considered here, the main couplings across components arise through shared system-level quantities, including the available budget system-level performance and risk constraints. Some of these shared variables can be provided to the decentralized actors, while the centralized critic uses the full system belief during training, allowing the actors to retain useful coordination signals without requiring high-dimensional full-system inputs that can make training slower and/or less stable. However, such localized actors may also be insufficient when assets are more strongly coupled through physical, functional, or network-flow dependencies. Recent work on the price of decentralization in MARL for infrastructure inspection and maintenance planning shows that decentralized methods can achieve near-optimal performance in series-like systems, whereas increasing redundancy can amplify coordination challenges and lead to optimality losses [Bhustali et al. \(2026\)](#). This suggests that the need for inter-agent information sharing can depend on the structure of system-level dependence and needs to be carefully examined before implementation. Further details on the DDMAC-CTDE architecture, including its implementation and performance, are provided in the following sections.

3.1. Formulation

A fully cooperative multi-agent task can be described as a Decentralized POMDP (Dec-POMDP) introduced in [\(Oliehoek & Amato, 2016\)](#). A Dec-POMDP consists of an environment with multiple agents, in which the agents do not have complete observations of the environment's state and the actions taken by the other agents. More formally, a Dec-POMDP is defined by a tuple $\langle \mathcal{N}_a, \mathcal{N}_c, \{S_i\}, \{A_i\}, \{\Omega_i\}, \mathbb{P}, \mathbb{O}, \mathbb{C}, \gamma \rangle$, where $\mathcal{N}_a$ is a set of agents from 1 to N_a , $\mathcal{N}_c$ is a set of components from 1 to N_c , S_k is a set of states for the k -th component and $S = S_1 \times \dots \times S_{N_c}$ is the global state of the system, A_i is a set of actions for agent i and $A = A_1 \times \dots \times A_{N_a}$ is the joint action space, Ω_i is a set of observations for the i -th agent and $\Omega = \Omega_1 \times \dots \times \Omega_{N_a}$ is the joint observation space, and $\mathbb{P}$, $\mathbb{O}$, and $\mathbb{C}$ are the same as defined earlier in [Section 2.1](#).

Without any loss of generality and for presentation simplicity, we henceforth assume that each agent controls exactly one corresponding system component, i.e., $N_a = N_c = N$. This assumption can easily be modified if an agent controls multiple components [Andriotis and Papakonstantinou \(2019a\)](#). Then, the policy $\pi_i : \mathbf{b}^{(i)} \rightarrow P(A_i)$ is a policy followed by the i -th agent and it is a function of the local belief state vector $\mathbf{b}^{(i)}$, determined by the observation and action history of each component and calculated via Bayesian update, as described in Eq. (5). For all agents, the joint policy $\pi = \pi_1 \cdot \pi_2 \cdot \dots \cdot \pi_{N-1} \cdot \pi_N$ may be defined as $\pi : \hat{\mathbf{b}} \rightarrow P(A)$, where $\hat{\mathbf{b}} = [\mathbf{b}^{(i)}]_{i=1}^N$ is a belief matrix of the system state. The joint objective of the agents is to identify an optimal policy π^* that minimizes the discounted cumulative global cost C_0 over the planning horizon T . For concise notation here, we are also using $(\cdot)'$ for $(\cdot)$ at the $t+1$ time step.

Similar to DDMAC, a decentralized actor-critic architecture is selected within the premises of off-policy actor critic DRL, where each agent is assigned an actor network for policy output independent to other agents. During training, the centralized critic is using the system state distribution as input and the value function $V^\pi(\hat{\mathbf{b}})$ as the output. The expression in Eq. (6) gives the optimal value function $V^{\pi^*}(\hat{\mathbf{b}})$. The actor network is updated using the policy gradient theorem [Sutton et al. \(2000\)](#) extended to multi-agent settings [\(Andriotis & Papakonstantinou, 2021\)](#). The off-policy gradient $\mathbf{g}_{\theta_\pi^{(j)}}$ for the j^{th} actor can be estimated using samples generated by a behavior policy $\mu \neq \pi$ as:

$$\begin{aligned} \mathbf{g}_{\theta_\pi^{(j)}} &= \mathbb{E}_{\mathbf{b} \sim \rho, \mathbf{a} \sim \mu} \left[w \left(\sum_{i=1}^N \nabla_{\theta_\pi^{(j)}} \log \pi_i \left(a^{(i)} | \mathbf{b}^{(i)}, \theta_\pi^{(i)} \right) \right) A^\pi(\hat{\mathbf{b}}, \mathbf{a}) \right] \\ &= \mathbb{E}_{\mathbf{b} \sim \rho, \mathbf{a} \sim \mu} \left[w \left(\nabla_{\theta_\pi^{(j)}} \log \pi_j \left(a^{(j)} | \mathbf{b}^{(j)}, \theta_\pi^{(j)} \right) \right) A^\pi(\hat{\mathbf{b}}, \mathbf{a}) \right] \end{aligned} \quad (7)$$

where $\mathbf{a} = \{a^{(i)}\}_{i=1}^N$ are the actions for N -agents, $\theta_\pi^{(j)}$ is a weight parameter for the j^{th} actor network, $w = \min\{\delta, \prod_{i=1}^N \pi_i(a^{(i)} | \mathbf{b}^{(i)}) / \mu_i(a^{(i)} | \mathbf{b}^{(i)})\}$ is a truncated importance sampling weight, μ is a N -dimensional vector of agents' behavior policies, ρ is the N -dimensional limiting state distribution under these policies, $\delta = 2$ is a cut-off value if the ratio becomes too large due to very small μ_i , and $A^\pi(\hat{\mathbf{b}}, \mathbf{a})$ is an advantage function defined as $A^\pi(\hat{\mathbf{b}}, \mathbf{a}) = Q^\pi(\hat{\mathbf{b}}, \mathbf{a}) - V^\pi(\hat{\mathbf{b}})$, hence being a zero mean measure, i.e., $\mathbb{E}_{\mathbf{a} \sim \pi} [A^\pi(\hat{\mathbf{b}}, \mathbf{a})] = 0$, estimating how advantageous the given action is. Advantage function can further be approximated as in [Andriotis and Papakonstantinou \(2019a\)](#); [Mnih et al. \(2016\)](#), similar to the TD-learning:

$$A^\pi(\hat{\mathbf{b}}, \mathbf{a} | \theta_V) \simeq c(\hat{\mathbf{b}}, \mathbf{a}) + \gamma V^\pi(\hat{\mathbf{b}}' | \theta_V) - V^\pi(\hat{\mathbf{b}} | \theta_V) \quad (8)$$

where θ_V are weight parameters of the value/critic network, and $\hat{\mathbf{b}}'$ is the system belief at the next time step. The mean squared error is considered as the loss function, L_V , for the critic network, given as:

$$L_V(\theta_V) = \mathbb{E}_{\mathbf{b} \sim \rho, \mathbf{a} \sim \mu} \left[w \left(c(\hat{\mathbf{b}}, \mathbf{a}) + \gamma V^\pi(\hat{\mathbf{b}}' | \theta_V) - V^\pi(\hat{\mathbf{b}} | \theta_V) \right)^2 \right] \quad (9)$$

and its gradient for updating θ_V can be estimated as:

$$\mathbf{g}_{\theta_V} = \mathbb{E}_{\mathbf{b} \sim \rho, \mathbf{a} \sim \mu} \left[w \nabla_{\theta_V} V^\pi(\hat{\mathbf{b}} | \theta_V) A^\pi(\hat{\mathbf{b}}, \mathbf{a} | \theta_V) \right] \quad (10)$$

Both actor and critic networks are trained using batch sampling from a buffer memory $\mathcal{M}$, containing experience tuples $(\mathbf{b}, \mathbf{a}, [\pi_i]_{i=1}^N, \mathbf{c}, \mathbf{b}')$.

4. Constrained optimization

The life-cycle management problem is often accompanied by several constraints, e.g., long or short-term budgets, risk tolerance, and other imposed restrictions by various agencies. Depending on their nature, constraints are handled in this work through state augmentation or Lagrangian multipliers.

4.1. Objective function with constraints

The objective of the decision-maker (agent) may be defined as determining an optimal policy $\pi = \pi^*$ that minimizes the entire cumulative operational costs and risks given several constraints [Andriotis and Papakonstantinou \(2021\)](#):

$$\begin{aligned} \pi^* &= \arg \min_{\pi \in \Pi} \mathbb{E}_{s_0:T, a_0:T, o_0:T} \left[\sum_{t=0}^T \gamma^t c \mid a_t \sim \pi(o_{0:t}, a_{0:t-1}), s_0 \sim \mathbf{b}_0 \right] \\ \text{s.t. } Z_{h,k} &= \sum_{t=0}^T \gamma_h^t z_{h,k}(s_t, a_t) - \alpha_{h,k} \leq 0, \quad k = 1, \dots, K \\ Z_{s,m} &= \mathbb{E}_{s_0:T, a_0:T, o_0:T} \left[\sum_{t=0}^T \gamma_s^t z_{s,m}(s_t, a_t, s_{t+1}) \right] - \alpha_{s,m} \leq 0, \quad m = 1, \dots, M \end{aligned} \quad (11)$$

where $c = c(s_t, a_t, s_{t+1})$ is the cost incurred at time t by taking action a_t , transitioning from state s_t to state s_{t+1} and taking an observation o_t ; γ , γ_h , and $\gamma_s \in [0, 1]$ are the discount factors for cost, hard constraints, and soft constraints, respectively; $\mathbf{b}_0$ is an initial belief; T is the length of the planning horizon; $Z_{h,k}$ and $Z_{s,m}$ are the *hard* and *soft* constraints, respectively; and $z_{h,k}$ and $z_{s,m}$ are their respective auxiliary costs; $\alpha_{h,k}$ and $\alpha_{s,m}$ are real-valued scalars, with the former determining the allowable values of K hard constraint thresholds that need to be satisfied at all time steps, and the latter representing the values for M soft constraint thresholds which need to be satisfied in expectation. More details on soft and hard constraints will be provided subsequently. The policy function belongs to a space $\pi \in \Pi_c$, which contains all possible policies that are admissible under the existing constraints of the problem. Π_c is a subset of Π , the policy space of the unconstrained problem.

The constraint structure in Eq. (11) is adaptable to a wide variety of constraint types important to infrastructure management. Hard constraints, for example, may be used to simulate a broad range of fixed resource allocation and control action availability problems, such as those with budget restrictions. In turn, soft constraints in the form of Eq. (11) may be used to represent constraints that need to be satisfied in expectation, such as risk-based or performance-based constraints (e.g., restricting poor condition infrastructure levels, imposing goals for good condition levels, and many others). Hard constraints can be handled through state augmentation, however, this is impractical for soft constraints since one must then track the full distribution of the cumulative discounted value of $z_{s,m}$. As a result, probabilistic constraints are handled using Lagrangian relaxation [Bertsekas \(1997\)](#). Based on the above, the optimization problem can be restated as in [Andriotis and Papakonstantinou \(2021\)](#):

$$\begin{aligned} V(\mathbf{b}_0, \mathbf{y}_0) &= \max_{\lambda_1, \dots, \lambda_M \geq 0} \min_{\pi \in \Pi} \mathbb{E}_{s_0:T, a_0:T, o_0:T} \left[\sum_{t=0}^T \left(\gamma^t \bar{c}(s_t, a_t, \mathbf{y}_t) \right. \right. \\ &\quad \left. \left. + \gamma_s^t \sum_{m=1}^M \lambda_m (z_{s,m} - \alpha_{s,m}) \right) \mid a_t \sim \pi(o_{0:t}, a_{0:t-1}, \mathbf{y}_t), s_0 \sim \mathbf{b}_0, \mathbf{y}_0 \right] \\ &= \max_{\lambda_1, \dots, \lambda_M \geq 0} \min_{\pi \in \Pi} V_\lambda^\pi(\mathbf{b}_0, \mathbf{y}_0), \end{aligned} \quad (12)$$

$$\begin{aligned} \text{s.t. } \mathbf{y}_t &= \{y_{kt}\}_{k=1}^K, \quad y_{kt} = \sum_{\tau=0}^{t-1} \gamma_h^\tau z_{h,k}(s_\tau, a_\tau), \quad y_{k0} = 0, \\ y_{kt} &\in [0, \alpha_{h,k}], \quad k = 1, \dots, K \end{aligned}$$

where variables y_{kt} track the discounted cumulative value of the function related to hard constraints, $Z_{h,k}$, up to time step $t-1$, and $\bar{c}(s_t, a_t, s_{t+1}, \mathbf{y}_t)$ is the total cost accumulated at a time step t taking $\mathbf{y}_t$ into consideration. Variables y_{kt} are upper bounded by $\alpha_{h,k}$. Lagrange multipliers, λ_m , constitute the dual variables of the max-min dual problem, they are positive scalars and are associated with the soft constraints.

Hard constraints:

Depending on the operational and resource allocation plan, available funds for inspection and maintenance must meet particular short- or long-term objectives associated with a certain budget cycle length, T_B . The total cost of inspection actions (c_I) and maintenance actions (c_M) at the time instant τ is given as follows:

$$z_h(s_\tau, a_\tau) = (c_M + \gamma c_I) \mathbf{1}_{\tau \in \Lambda_t} \quad (13)$$

where $\Lambda_t = ([t/T_B] T_B, ([t/T_B] + 1) T_B]$ is a budget interval of length T_B , $[x]$ is the integer part of x . For a given budget cap α_h , the maintenance and inspection costs at each time are adjusted as:

$$\begin{aligned} \bar{c}_M &= \mathbf{1}_{y+z_h \leq \alpha_h} c_M \\ \bar{c}_I &= \mathbf{1}_{y+z_h \leq \alpha_h} c_I \end{aligned} \quad (14)$$

Eqs. (13) and (14) essentially affect inspection and maintenance costs, such that the accumulated cost $z_h(s_\tau, a_\tau)$ in a given budget interval T_B does not exceed the budget limit α_h for the interval. Introducing a new state variable $y = y_t$ allows for the inclusion of budget constraints, which are assumed to be fully observable. This enables the agent to justify control actions based on the available budget, $\alpha_h - y_t$, at each time step.

Soft constraints:

Soft constraints are the constraints that need to be satisfied in expectation over the life-cycle of the considered system. The generic formulation provided in [Andriotis and Papakonstantinou \(2021\)](#) is as follows:

$$\begin{aligned} Z_s &= \mathbb{E}_{s_0:T, a_0:T, o_0:T} \left[\sum_{t=0}^T \gamma_s^t J_t^\pi \right] - \alpha_s \\ &= J^\pi - \alpha_s \end{aligned} \quad (15)$$

where Z_s is a constraint function, J^π is a cumulative discounted cost (or any generic parameter that needs to be bounded) accrued by following policy π , J_t^π is a cost (or the parameter value) at time step t , and α_s is the cap that needs to be satisfied over the life-cycle.

Many reliability-related decision problems are focused on limiting the likelihood of failure (i.e., the probability of reaching a failure state from a non-failure state) throughout the system's operational life. For transportation network management, it is common to limit the poor condition states of pavement and bridge components based on state and federal criteria and thresholds. This study will demonstrate the incorporation of soft constraints in such cases. The details of the performance criteria will be discussed in [Section 7.4](#).

4.2. Costs related to infrastructure management

Costs at different time steps for a selected action can be decomposed into inspection cost, $c_I(a_{I,t})$, maintenance cost, $c_M(a_{M,t})$, and user delay cost, $c_D(a_{M,t})$, incurred in this case due to the traffic delays from maintenance actions, and risk cost $c_{\text{risk}}(s_t, a_t, s_{t+1})$. In addition, other costs could be considered, such as vehicle operating costs possibly incurred due to

the system's current state, for example, extra fuel and tire costs due to poor ride quality, environmental costs due to fuel emissions which will also depend on the system's state, and various other costs. Therefore, the total cost at each decision step can be generally expressed as:

$$c(s_t, a_t, s_{t+1}) = c_M(a_{M,t}) + c_D(a_{M,t}) + c_{\mathcal{R}\pi}(s_t, a_t, s_{t+1}) + \gamma c_I(a_{I,t}) \quad (16)$$

where $a_{I,t} \in A_I$ is the selected inspection and $a_{M,t} \in A_M$ is the selected maintenance action at time step t . Under this distinctive consideration of actions, the total action can be defined as $a_t \in A = A_I \times A_M$. For finite horizon problems, often a terminal cost (c_{s_T}) is added to guide actions at the end of the horizon, i.e., at time step $t = T$. After incorporating the budget constraint at time t and c_{s_T} at time step T , the cost can be modified as:

$$\begin{aligned} \bar{c}(s_t, a_t, s_{t+1}, \mathbf{y}_t) &= \bar{c}_M(a_{M,t}) + \gamma \bar{c}_I(a_{I,t}) + \bar{c}_D(a_{M,t}) \\ &+ \bar{c}_{\mathcal{R}\pi}(s_t, a_t, s_{t+1}) + \mathbf{1}_{t=T} \bar{c}_{s_T}(s_t) \end{aligned} \quad (17)$$

The transition and observation probabilities are also affected by the budget constraint, because after utilizing the budget in a given budget cycle, the agents can only choose actions with no costs, as no maintenance and inspection actions can be performed. Further, the costs in Eqs. (16) and (17) can readily be obtained by calculating cost in expectation with respect to the belief state $\mathbf{b}$ as in (Andriotis & Papakonstantinou, 2021):

$$\begin{aligned} \bar{c}_{b,t} &= \bar{c}_{b,t}(\mathbf{b}_t, \mathbf{a}_t, \mathbf{y}_t) \\ &= \mathbb{E}_{s_t, s_{t+1}} [\bar{c}(s_t, a_t, s_{t+1}, \mathbf{y}_t)] \\ &= \sum_{s_t \in \mathcal{S}} b(s_t) \sum_{s_{t+1} \in \mathcal{S}} P(s_{t+1} | s_t, a) \bar{c}(s_t, a_t, s_{t+1}, \mathbf{y}_t) \end{aligned} \quad (18)$$

where $\bar{c}_{b,t}$ is the estimated total cost at time t , obtained using belief state $\mathbf{b}_t$.

4.3. DDMAC-CTDE formulation with constraints

The architecture given in Section 3 can readily be extended to a constrained case. Eqs. (7)-(10) are modified to incorporate Eqs. (12)-(18). Similarly, here we also have N components and N number of agents, where one agent has control over exactly one component. For such settings, the system policy function at timestep t can be written as:

$$\pi(\mathbf{a}_t | \hat{\mathbf{b}}_t, \mathbf{y}_t, t) = \prod_{i=1}^N \pi_i(a_t^{(i)} | \mathbf{b}_t^{(i)}, \mathbf{y}_t, t) \quad (19)$$

where $\mathbf{y}_t = [\mathbf{y}_t^k]_{k=1}^K$ denotes the augmented state vector containing the current values of the *hard constraints* at time step t . The input to actor i is thus represented here by its local component belief state, the shared system-level hard-constraint or budget state, and the current time step. The complete input specifications are provided in Section 8.1, while a simplified notation is retained here for clarity. Decentralized execution in this work thus means that each actor does not require the full system belief or the beliefs of some other components at deployment, while still using shared public information, such as the remaining budget and planning year. The decentralized policy given in Eq. (19) and the centralized Lagrangian value function can be parameterized as:

$$\begin{aligned} \pi_i(a_t^{(i)} | \mathbf{b}_t^{(i)}, \mathbf{y}_t, t) &\simeq \pi_i(a_t^{(i)} | \mathbf{b}_t^{(i)}, \mathbf{y}_t, t, \theta_\pi^{(i)}) \\ V_\lambda^\pi(\hat{\mathbf{b}}_t, \mathbf{y}_t, t) &\simeq V_\lambda^\pi(\hat{\mathbf{b}}_t, \mathbf{y}_t, t | \theta_V) \end{aligned} \quad (20)$$

where $\theta_\pi^{(i)}$, θ_V are weight parameters for policy and value network, respectively. Training is performed off-policy with experience replay $\mathcal{M}$. The replay buffer contains agents' collected experiences at timesteps t as tuples of the form $(\hat{\mathbf{b}}_t, \mathbf{y}_t, \mathbf{a}, [\pi_i]_{i=1}^N, \bar{c}_b, [z_{s,m}]_{m=1}^M, \hat{\mathbf{b}}', \mathbf{y}')$, where the primed variables correspond to the quantities at the next timestep, $t + 1$. $\bar{c}_b$

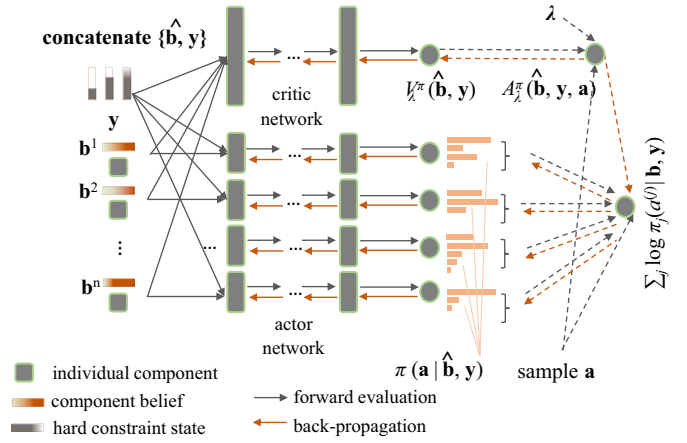


Fig. 1. Constrained Deep Decentralized Multi-Agent Actor-Critic (DDMAC) architecture under the centralized training and decentralized execution (CTDE) paradigm, illustrating the inputs to the decentralized actor and centralized critic networks at time step t .

is a cost collected based on the beliefs instead of states. The modified off-policy gradients are now given:

$$\nabla_{\theta_\pi^{(j)}} V_\lambda^\pi = \mathbb{E}_{\mathcal{M}} \left[w \left(\sum_{i=1}^N \nabla_{\theta_\pi^{(i)}} \log \pi_i(a_t^{(i)} | \mathbf{b}_t^{(i)}, \mathbf{y}_t, t, \theta_\pi^{(i)}) \right) A_\lambda^\pi(\hat{\mathbf{b}}_t, \mathbf{y}_t, \mathbf{a}_t) \right] \quad (21)$$

$$\nabla_{\theta_V} V_\lambda^\pi = \mathbb{E}_{\mathcal{M}} \left[w \nabla_{\theta_V} V_\lambda^\pi(\hat{\mathbf{b}}_t, \mathbf{y}_t, t | \theta_V) A_\lambda^\pi(\hat{\mathbf{b}}_t, \mathbf{y}_t, \mathbf{a}_t) \right] \quad (22)$$

$$\nabla_{\lambda_m} V_\lambda^\pi \simeq \sum_{t=0}^T \gamma^t z_{s,m} - \alpha_{s,m} \quad (23)$$

where λ_m is a Lagrange multiplier and A_λ^π is an advantage function, computed here as:

$$A_\lambda^\pi(\hat{\mathbf{b}}_t, \mathbf{y}_t, \mathbf{a}_t | \theta_V) \simeq \bar{c}_b + \sum_{m=1}^M \lambda_m z_{s,m} + \gamma V_\lambda^\pi(\hat{\mathbf{b}}', \mathbf{y}', t+1 | \theta_V) - V_\lambda^\pi(\hat{\mathbf{b}}_t, \mathbf{y}_t, t | \theta_V) \quad (24)$$

The full DDMAC-CTDE with constraints schematic is presented in Fig. 1, and the implementation steps are described in Algorithm 1. In the subsequent sections the specific modeling of pavements, bridges and the network system is described.

5. Pavement modeling

There are various indicators to characterize the condition state of pavements, such as the Pavement Condition Index (PCI), Critical Condition Index (CCI), International Roughness Index (IRI), and Load Related Distress Index (LDR), among many others Hasnat et al. (2023). CCI and IRI are used in this work, consistent with the Virginia State Pavement Management System.

5.1. Critical condition index (CCI)

The CCI model in this study is a modified version of the modeling described in VDOT report Katicha et al. (2016), where a regression model is presented that estimates mean CCI, using collected data from the years 2007-12, and 2014-15, based on the Virginia Department of Transportation (VDOT) Pavement Management System (PMS). The age of the pavement is calculated as the difference between the year of condition reporting and the last year of relevant recorded maintenance. The total dataset in Katicha et al. (2016) consists of 3,473 observations from the years 2007 to 2012, and 1,560 observations from the years 2014 and 2015. The mean regression model in Katicha et

Algorithm 1: DDMAC-CTDE with constraints.

```

Initialize replay buffer (empty);
Initialize actor, critic network weights  $[\theta_\pi^{(j)}]_{j=1}^N, \theta_V$  randomly,
and Lagrange multipliers  $[\lambda_m]_{m=1}^M$  with zeros;
for episode = 1,  $N_{ep}$  do
  for  $t = 1, T$  do
    Select action  $\mathbf{a}_t$  at random according to exploration noise;
    Otherwise select action  $\mathbf{a}_t \sim \mu_t = [\pi_j(\cdot | \mathbf{b}_t^{(j)}, \mathbf{y}_t, t; \theta_\pi^{(j)})]_j^N$ ;
    Estimate costs  $\bar{c}_{b,t}$  as:
    if  $t \neq T$  then
       $\bar{c}_{b,t} = \bar{c}_b$ ;
    else
       $\bar{c}_{b,t} = \bar{c}_b + \bar{c}_{s,T}$ ;
    calculate  $z_{s,mt} = z_{s,m}$  given  $\hat{\mathbf{b}}_t$  and  $\mathbf{a}_t$ ;
    Observe  $o_{t+1}^{(l)} \sim p(o_{t+1}^{(l)} | \mathbf{b}_t^{(l)}, \mathbf{a}_t)$  for  $l = 1, 2, \dots, N$ ;
    Compute beliefs  $\mathbf{b}_{t+1}^{(l)}$  for  $l = 1, 2, \dots, N$ ;
    Store experience  $(\hat{\mathbf{b}}_t, \mathbf{y}_t, \mathbf{a}_t, \mu_t, \bar{c}_{b,t}, [z_{s,mt}]_{m=1}^M, \hat{\mathbf{b}}_{t+1}, \mathbf{y}_{t+1})$  to
    replay buffer;
  end
  Sample batch of  $(\hat{\mathbf{b}}_t, \mathbf{y}_t, \mathbf{a}_t, \mu_t, \bar{c}_{b,t}, [z_{s,mt}]_{m=1}^M, \hat{\mathbf{b}}_{t+1}, \mathbf{y}_{t+1})$  from
  the buffer;
  If  $\hat{\mathbf{b}}_t$  is terminal state  $A_t = \bar{c}_{b,t} + \sum_{m=1}^M \lambda_m z_{s,mt} - V_\lambda^\pi(\hat{\mathbf{b}}_t, \mathbf{y}_t, t | \theta_V)$ ;
  Otherwise:  $A_{\lambda,i}^\pi =$ 
     $\bar{c}_{b,t} + \sum_{m=1}^M \lambda_m z_{s,mt} + \gamma V_\lambda^\pi(\hat{\mathbf{b}}_{t+1}, \mathbf{y}_{t+1}, t+1 | \theta_V) - V_\lambda^\pi(\hat{\mathbf{b}}_t, \mathbf{y}_t, t | \theta_V)$ ;
  Update actor parameters  $\theta_\pi^{(j)}$  according to gradient:
     $\nabla_{\theta_\pi^{(j)}} V_\lambda^\pi \simeq \sum_i w_i \left( \sum_{u=1}^N \nabla_{\theta_\pi^{(j)}} \log \pi_u \left( a_i^{(u)} | \mathbf{b}_t^{(u)}, \mathbf{y}_t, i; \theta_\pi^{(u)} \right) \right) A_{\lambda,i}^\pi$ ;
  Update critic parameters  $\theta_V$  according to gradient:
     $\nabla_{\theta_V} V_\lambda^\pi \simeq \sum_i w_i \nabla_{\theta_V} V_\lambda^\pi(\hat{\mathbf{b}}_t, \mathbf{y}_t, t | \theta_V) A_{\lambda,i}^\pi$ ;
  Update Lagrange multipliers  $\lambda_m, m = 1, \dots, M$ , according to
  gradient:  $\nabla_{\lambda_m} V_\lambda^\pi \simeq \sum_{t=0}^T \gamma^t z_{s,mt} - \alpha_{s,m}$ ;
end

```

al. (2016) directly incorporates the effect of the age of the pavement section and indirectly considers the impact of the traffic load and other pavement structural parameters. Hence, based on this modeling, the effective age of the pavement section also determines its deterioration rate over time. Fig. 2 shows the mean CCI as a function of pavement age for different traffic levels (A to E), where level A to E indicates heavy to light traffic conditions, assuming other structural parameters are constant. In this context, while maintaining general applicability and without any loss of generality, we make the assumption that the depicted curves for various pavements serve as indicative examples for certain pavement design criteria and traffic levels. It is important to note that for alternative design criteria and traffic levels, the resulting curves may exhibit variations. Here, we have thus illustratively assigned deterioration curve A to the representative interstate pavement type, curve C to the primary pavement type, and curve E to the secondary pavement type. Then, a nonstationary gamma process is fitted in the following section to estimate the relevant nonstationary transition probabilities.

5.1.1. Fitting a nonstationary gamma process

A gamma process is utilized in this section, with its mean in time equal to the modified mean CCI predictions in Fig. 2 and a relevant model variance $\sigma_s^2(t)$, primarily determined based on Katicha et al. (2016) and the observed data trends therein. The marginal probability of

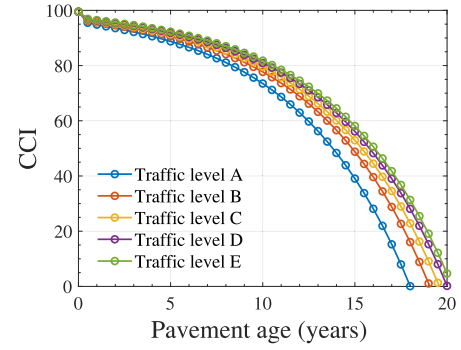


Fig. 2. Modeled mean CCI for different levels of traffic.

CCI at every step t can be obtained by estimating the damage index DI , where $DI = 100 - CCI$, using the gamma distribution with probability density function:

$$Ga(DI|f(t), g(t)) = \frac{g(t)^{f(t)}}{\Gamma(f(t))} DI^{f(t)-1} e^{-g(t)DI} \quad (25)$$

The gamma process is parameterized here by a non-negative time varying scale parameter function $g(t)$ and a non-negative time-varying shape parameter function $f(t)$, and $\Gamma(u) = \int_0^\infty v^{u-1} e^{-v} dv$. The relevant parameters are estimated in time based on the mean DI prediction model and the model variance $\sigma_s^2(t)$. The relationships between the gamma process parameter functions and μ_{DI} , σ_s are given as:

$$\mu_{DI}(t) = \frac{f(t)}{g(t)}, \sigma_s(t) = \frac{\sqrt{f(t)}}{g(t)} \quad (26)$$

Due to their monotonicity, gamma processes are readily used as a suitable modeling choice in stochastic deterioration engineering applications and can describe continuous Markovian transitions. For time instant $t_1 < t_2$, the increment of DI follows a gamma distribution:

$$DI(t_2) - DI(t_1) \sim Ga(\cdot | f(t_2) - f(t_1), g(t_2)) \quad (27)$$

where $g(t)$ is assumed to be constant in the interval $(t_1, t_2]$. In Fig. 3, relevant simulation results are indicatively shown for traffic level A with 300 different realizations. All corresponding $f(t)$ and $g(t)$ values are shown in Table C.5 in the Appendix, for all different traffic levels.

The nonstationary gamma process used here does not explicitly represent spatial correlation among component deterioration processes. This conditional independence assumption is adopted to adhere to currently used, common, practical asset-management modeling approaches, where component-wise deterioration models are used, due to differences in materials, construction histories, maintenance actions, and localized damage mechanisms. In general, spatial dependence may still arise though, for example through shared traffic loading, environmental exposure, drainage conditions, or other regional factors, and, if desired, could be incorporated into our framework Morato et al. (2023). The developed DDMAC-CTDE approach could still be applied without any limitations to such spatially correlated environments. However, the underlying inference problem becomes more complex, as observations from one component could update beliefs about other correlated components, requiring joint or structured belief representations Morato et al. (2023).

5.1.2. Determining transition probabilities

To calculate the transition probabilities, one must first specify the discrete condition states, as given in Table 1. The discretized condition states in table are mostly based on the VDOT's prescribed maintenance

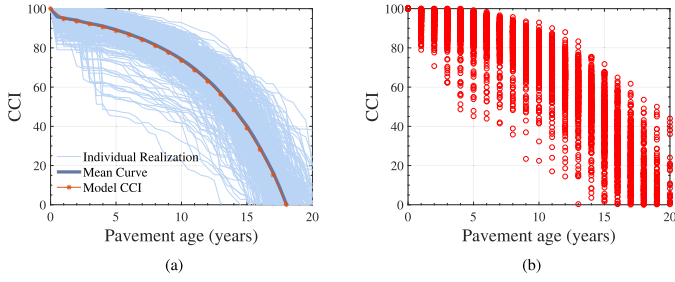


Fig. 3. (a) Fitted gamma model, (b) Scatter plot for CCI for traffic level A.

Table 1

State discretization based on CCI values.

CCI state (s)	CCI values	Pavement Condition
$s = 6$	100–90	Excellent
$s = 5$	89–80	Very Good
$s = 4$	79–61	Good
$s = 3$	60–50	Fair
$s = 2$	49–37	Poor
$s = 1$	<37	Very Poor

recommendations in Chowdhury (2016), which are as follows for interstate highways: For CCI values greater than 89, the treatment category is always Do Nothing (DN); for values greater than 84, the treatment category is always DN or Preventive Maintenance (PM); for values between 60 and 84, the treatment category is PM; for values less than 60, the treatment category is Corrective Maintenance (CM), Restorative Maintenance (RM), or Rehabilitation/Reconstruction (RC); for values less than 49, the treatment category is always RM or RC; and for values less than 37, the treatment category is always RC. The detailed definitions of VDOT maintenance actions are provided in Table E.1 in the Appendix.

Monte Carlo sampling is then performed by generating CCI and pavement age pairs, propagated in time following the described gamma process. A million sequences are generated in total to obtain the transition probabilities for a given traffic level. Fig. 4 indicatively shows computed transition probabilities for traffic level A. Various advanced sampling techniques, like Hamiltonian Monte Carlo Arcieri et al. (2023); Papakonstantinou et al. (2023a), can naturally be used for more challenging distributions.

5.1.3. Observation probabilities for CCI inspection action

Observation uncertainty can be appropriately modeled by the likelihood functions $p(o_t|s_t)$ which quantify the probability of observing an observation o_t at time t given a state s_t at that instant. To calculate the observation probability $p(o_t|s_t)$ a normal distribution is considered with mean equal to the actual CCI value and an error variance $\sigma_{error}^2 = 72$, based on Katicha et al. (2016), and is considered as a low fidelity

inspection due to the significant observation uncertainty. The error variance is linked to the variability in the measurement of CCI and the error in the reported pavement condition by an inspector. These conditional observation probabilities, $p(o_t|s_t)$, are generated by calculating the area under the curve of the normal distribution.

It has been observed that the observation probabilities are largely independent of different traffic levels, so the average values have been reported in Table E.2, which can be used for all traffic level values. High-fidelity inspections can also be considered, which rely on higher precision, expensive devices, and multi-modality. These techniques may also require significant post-processing which increases the overall inspection cost. Hence, a much smaller error variance, $\sigma_{error}^2 = 18$, is used to compute the observation probabilities, as compared to the previous case. Using these inspection techniques, we derive the observation probabilities as shown in Table E.3. The no-inspection case is also mathematically equivalent to when one has an equal probability of observing each state i.e., $p(o_t|s_t) = 1/6 \forall o_t, s_t \in \{1, 2, \dots, 6\}$. In Table E.2, we notice observation probabilities less than 50% for observing the actual states, which are attributed here to the non-uniform CCI partitioning, as mentioned in Section 5.1.2.

As discussed earlier, the cost of inspections depends on the techniques used for observing the pavement condition along with the post-processing time/complexity. The inspection techniques have been categorized into three precision categories of no-inspection, low-fidelity inspection, and high-fidelity inspection, as previously mentioned, and the associated costs are informed by the Michigan DOT survey Dye Management Group Inc (2014) for monitoring highway assets. These costs are summarized in Table E.10 in the Appendix.

5.1.4. Maintenance, rehabilitation, and reconstruction actions in relation to CCI

For ease of notation and to simplify expressions, the word maintenance is used henceforth as a general term encompassing all actions in relation to preventive treatments, repairs, restorations, and reconstructions. There are various guidelines for pavement maintenance for different agencies/states. Along similar lines to VDOT Chowdhury (2016), four different maintenance actions are utilized here, i.e., (*Do Nothing*, *Minor Repair*, *Major Repair*, and *Reconstruction*). *Do Nothing* is the same as DN by VDOT, and *Minor Repair* is more aligned with VDOT's Preventive Maintenance (PM) category. *Major Repair* is, accordingly, similar to Corrective Maintenance (CM) and Restorative Maintenance (RM), and lastly *Reconstruction* is the same as VDOT's Reconstruction action. Further details related to VDOT's maintenance actions are provided in the Appendix Table E.1.

The maintenance action transition probabilities for CCI are shown in Tables E.4–E.6 similar to the line of work presented in Madanat and Ben-Akiva (1994). It is assumed that the *Minor Repair* (crack filling, moderate patching, etc.) cannot change the rate of deterioration, or the effective age of the component, but only its condition state. The change in condition states is considered based on Table E.4. For a *Major Repair* maintenance action, the deterioration rate is however reduced to that of an asset younger by 5 years or to that of a newly built pavement component, i.e., $(T_{age})_{updated} = \max(0, T_{age} - 5)$ years. *Major Repairs* also alter the CCI conditions, with the transition probabilities shown in Table E.5. The *Reconstruction* maintenance action brings the CCI condition back to the intact state (i.e., $s_t = 6$) with certainty, as shown in Table E.6, and alters the deterioration rate of the component to the initial intact value, or equivalently assigns it an effective age of 0. Transitions in these tables do not include the transition due to the environment and its deterioration effects. Decoupled environmental transitions can essentially allow us to include action duration and probability effects of successfully completed actions, among others. In Table E.4, we can observe that after applying maintenance actions the system can still transition to a worse state with a non-zero probability. This happens because we are accounting for the inherent uncertainty present in the outcome of the applied *Minor Repair* action, which also includes the possibility of failed actions.

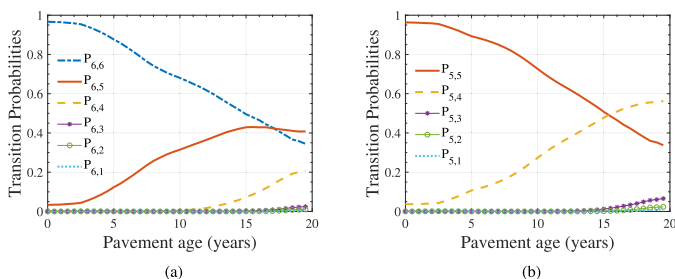


Fig. 4. Transition probabilities for traffic level A, with (a) starting state = 6, (b) starting state = 5, smoothed over time.

Table 2
Maintenance action costs for asphalt pavements, reported in USD/m².

Actions	Description	Cost (USD/m ²) Interstate	Cost (USD/m ²) Primary	Cost (USD/m ²) Secondary
Do Nothing	NA	0.00	0.00	0.00
Minor Repair	Moderate patching (<10%), surface treatment, partial depth patching, thin Asphalt Concrete (AC) overlay	20	16	10
Major Repair	Heavy patching (<20% of the pavement area), full depth patching, structural overlay	75	68	52
Reconstruction	Replacement of the entire pavement section	350	330	250

Table 3
State classification based on IRI (m/km) values.

Condition State (s)	IRI (m/km)	Pavement Condition
s = 5	< 0.95	Very Good
s = 4	0.95-1.56	Good
s = 3	1.57-2.19	Fair
s = 2	2.20-3.14	Mediocre
s = 1	> 3.15	Poor

Table 4
Do-Nothing transition probabilities for 5 IRI states of asphalt pavements.

Condition State	$s_{t+1} = 5$	$s_{t+1} = 4$	$s_{t+1} = 3$	$s_{t+1} = 2$	$s_{t+1} = 1$
$s_t = 5$	0.840	0.121	0.039		
$s_t = 4$		0.788	0.142	0.070	
$s_t = 3$			0.708	0.192	0.01
$s_t = 2$				0.578	0.422
$s_t = 1$					1.000

5.1.5. Maintenance action costs and durations

The cost of different maintenance actions is considered from American Road and Transportation Builders Association (2026); Chowdhury (2016); Cost per mile models by FDOT (2022); PennDOT (2017) and provided for interstate, primary, and secondary roads in Table 2. It is reported in USD/m² and the total cost can be calculated using the components' length and width (which is considered as 3.7 m or 12 ft per lane). Here, it is important to mention that the action descriptions provided in Table 2 for *Minor Repair* and *Major Repair* are indicative actions of the corresponding category and these actions are a subset of actions used in Chowdhury (2016), as presented in the Appendix in Table E.1.

The duration of different maintenance actions is inferred from ADOT (2018); PennDOT (2019). It is further assumed that no action takes more than 2 years, of which only *Reconstruction* actions can take more than one year to complete for a component. This can be justified by assuming multiple maintenance activities at a time for longer and multi-lane components. These durations are provided in Table E.12. If a *Reconstruction* action taking more than a year is performed by the agent, the component will not be available for the next time-step as part of the active component set. Thus, no action is recommended for this component at the next time-step, and this can be tracked in our DRL framework using an additional state.

5.2. International roughness index (IRI)

IRI (m/km) is another metric that quantifies the functional pavement condition in terms of the roughness experienced by vehicle passengers. IRI is calculated from measured longitudinal road profiles using the simulated vertical displacement response of a quarter-car Gillespie et al. (1986). IRI (m/km) in this work is discretized into 5 states, as in Faddoul et al. (2013); United States Department of Transportation (2000) presented in Table 3. The transition probabilities for the *Do Nothing* action are now given in Table 4.

5.2.1. Maintenance actions in relation to IRI

The same four maintenance actions are considered for IRI: *Do Nothing*, *Minor Repair*, *Major Repair*, and *Reconstruction*. The IRI transitions for *Minor Repair* to *Reconstruction* actions are given in Tables E.7–E.9 respectively, similar to the CCI case. Furthermore, the cost of maintenance actions in relation to IRI is the same as the one reported in Table 2. Maintenance actions taken at any given time will improve both CCI and IRI indices simultaneously.

Table 5
Pavement inspection cost (i.e., combined cost of IRI and CCI).

Inspection Technique	Description	Cost (in USD/m ²)
i_2	High fidelity	0.20
i_1	Low fidelity	0.10
i_0	No inspection	0.00

5.2.2. Observation probabilities for IRI inspection actions

Three assumed inspection actions of different fidelities are again considered in this case, and the measurement errors related to the respective inspection technologies are assumed to be normally distributed with zero mean and standard deviations (SD) of 0.08, 0.32, and ∞ m/km, respectively, as reported in Faddoul et al. (2013). The observation probabilities are summarized in Table E.11, where o_t is the observed state and s_t is the (hidden) actual state. The underlying assumption is that more costly inspections are high-fidelity inspections that provide more accurate information. For the last inspection technique, an infinite standard deviation is considered. Thus, this inspection is completely uninformative, equivalent to a no-inspection action.

Inspection costs are considered independent of the pavement state, as also considered in Faddoul et al. (2013), and depend on the fidelity of the inspection technology that is utilized. In Table E.11 costs are equal to 0.10, 0.03, 0 USD per m², respectively. Table E.10 and Table E.11 provide the costs of individual inspection actions for CCI and IRI, respectively, while in Table 5 the joint cost of pavement inspection is considered when both IRI and CCI are observed, as adopted in this work when an inspection action is chosen.

6. Bridge modeling

Decks are considered in this study to characterize the bridge condition and inform action selection, due to their expensive and frequent maintenance needs, in relation to other bridge subsystems. Decks often have the most intense rate of deterioration, among bridge components, with factors such as corrosion, cracking, freezing and thawing cycles, and direct exposure to traffic loads contributing to their degradation. Their condition becomes the dominant factor in determining the overall bridge condition rating.

6.1. Bridge decks

For bridge deck condition, 9 total states are generally considered, with state 9 being the undamaged state, as per the convention adopted

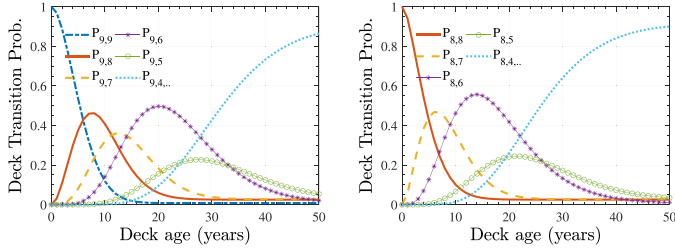


Fig. 5. Transition probabilities over time, moving from state 9 (left) and state 8 (right) to lower states.

Table 6

Bridge failure probability given bridge deck condition state.

Condition State	$s_t = 9$	$s_t = 8$	$s_t = 7$	$s_t = 6$	$s_t = 5$	$s_t = 4, \dots$	$s_t = \text{failed}$
Failure prob. ($s_{t+1} = \text{failed}$)	0.001	0.001	0.005	0.005	0.005	0.01	1.000

by PennDOT Pennsylvania Department of Transportation and Bureau of Design (2022) and other DOTs. For this report, state 4, ... characterizes all subsequent states, as also considered in Manafpour et al. (2018). Thus, 6 bridge deck states along with a failure state are used. The transition probabilities related to these 6 condition states are based on approximately 30 years of in-service performance data for over 22,000 bridges in Pennsylvania, as analyzed in Manafpour et al. (2018) for different exogenous parameters, i.e., deck length, location, protective coating, span type, etc. Fig. 5 shows an indicative case based on the following exogenous parameters: Deck length = 90 ft; Location (district) type = 4; Reinforcement bar protection = 1 (no protective coating); Type of span interaction for main unit = 1 (simple, non-composite); Primary load carrying members = 1 (reinforced concrete); Deck surface type = 1 (concrete); Total number of spans in main unit = 2 (multi-span); NHS = 1 (interstate route); Maintenance type = type 1 (no maintenance). Apart from these 6 nonstationary transitions, stationary failure probabilities are also considered, as shown in Table 6. Note that these failure probabilities are indicative; the actual failure probabilities depending on multiple factors, such as bridge type, location, hazard and other factors, can be computed using methods presented in Papakonstantinou et al. (2023b,2023a).

6.1.1. Maintenance actions for bridge decks

Similar to pavements, four maintenance actions are considered for maintaining the bridge decks: *Do Nothing*, *Minor Repair*, *Major Repair*, and *Reconstruction*. It is again assumed here that the *Minor Repair* action does not change the rate of deterioration of the deck, but it can change the condition state of the structure, as per the transition probabilities given in Table E.13. A *Major Repair* action, however, can improve the deterioration rate to that of a deck younger by five years or reset it to a newly made deck, i.e., $(T_{age})_{updated} = \max(0, T_{age} - 5)$ years, and the change in condition state is now considered as reported in Table E.14. The transition probabilities for the *Reconstruction* action are shown in Table E.15 and reset the state and deterioration rate of the deck to a newly built one.

6.1.2. Maintenance action costs and durations for bridge decks

The cost of maintenance actions has a small dependence upon the current state of the bridge deck. As a result, maintenance costs are considered independent of the state of the deck here. The cost for a reconstruction action is obtained from Federal Highway Administration (2019) and the costs for minor and major maintenance actions are considered as 15% and 45% of that, respectively, as inferred from Wells (1995). Table 7 shows the costs of performing individual actions in USD/m². It can be observed that the unit costs of actions for bridges are higher than the unit costs of pavement actions.

Table 7

Cost of maintenance actions for bridge decks in USD/m².

Actions	Description	Cost (USD/m ²)
Do Nothing	NA	0.00
Minor Repair	Moderate cracks filling and patching area < 10% of the deck area, minor replacement of reinforcement	400.00
Major Repair	Fixing spalls/delamination with deck area < 25%, major replacement of reinforcement	1,200.00
Reconstruction	Replacement of the entire deck	2,650.00

Table 8

Inspection action costs for bridge decks for 3 fidelity levels.

Inspection Technique	Description	Cost (USD/m ²)
i_2	High fidelity	1.20
i_1	Low fidelity	0.50
i_0	No Inspection	0.00

The duration of different maintenance actions is mainly inferred from ADOT (2018); Goodspeed and Brown (2017); Oakgrove (2013). For large bridges, it is assumed that maintenance activity is performed at multiple locations simultaneously to increase efficiency. The action durations are provided in Table E.18 for 3 different types of bridges, types I-III, categorized based on their sizes. Type I are the largest size bridges with heavy traffic, thus, it takes the longest to repair. Type II are intermediate size bridges with medium traffic, thus it takes less time than type I bridges to repair. Type III are small size bridges with the lowest traffic, needing the least amount of time to repair. More details about different types of bridges are provided in Section 7.1.

6.1.3. Inspection actions and observation probabilities

There is a variety of destructive and non-destructive inspection techniques that are used for bridge decks, such as visual inspection, acoustic sensing, infrared/thermal imaging, ground penetrating radar, coring and chipping, and half-cell potential tests, among many others. Based on their accuracy, inspection techniques can be characterized as i_0 , i_1 , and i_2 corresponding to uninformative, low-fidelity, and high-fidelity inspections, respectively, and the relevant inspection costs are related to their accuracy. As assumed for pavements, we here again relate the inspection costs to the reconstruction cost of the bridge and we consider 0.1% and 0.5% of the rebuild cost for low- and high-fidelity inspection techniques, respectively. The resulting costs are shown in Table 8.

The observation probabilities for the bridge decks are considered similarly to Faddoul et al. (2013); Madanat and Ben-Akiva (1994); Madanat (1993) and for the no-inspection case $p(o_t | s_t) = 1/6 \forall o_t, s_t \in \{1, 2, \dots, 6\}$ is used, as the failure state is assumed to be self announcing. Table E.16 and Table E.17 describe the observation probabilities for low- and high-fidelity inspection techniques, respectively.

7. Transportation network

7.1. Topology and functionality

As a testbed, we consider a modified version of the Hampton Roads transportation network, in Virginia, USA. The original topology, road characteristics, and daily traffic data of the network are used, together with 11 main network bridges for this cross-asset environment, that are related to important network routes. Each bridge is bidirectional, with the same number of lanes as the original network, as illustrated in Fig. 6. Different decks, types I-III, are categorized based on their relative sizes and traffic volumes in descending order. Type I bridges are large bridges, typically having a length of more than 2 miles, with the highest volume

of traffic; type II bridges range between 0.4- and 2 miles with intermediate traffic volumes; and type III bridges usually have lengths less than 0.4 miles and have the lowest traffic volumes. An exception to this classification is the Great Bridge bypass, which should be of type II based on its size, but as it is on a secondary highway with no nearby interstates or high-traffic volume highways, it is categorized as type III. More details on bridge characteristics are shown in Table B.2. Traffic volumes, which can be inferred from *Daily Vehicle Miles Traveled (DVMT)* (2020), are also related to user delay costs associated with maintenance actions, as also explained later in more detail.

Similarly, the network has various pavement components categorized as Type I-III. Type I pavements are associated with interstate highways with bidirectional traffic having four lanes in each direction, thus, constituting the highest vehicular mileage per mile. Type II pavements correspond to the primary state highways with bidirectional medium-level traffic, having two lanes in each direction. Lastly, Type III pavements characterize the secondary state highways with low-level bidirectional traffic, with one lane in each direction. Fig. 6 (left) shows the Hampton Roads map, and Fig. 6 (right) represents the schematic of the network with green links pertaining to primary and secondary highways and red links representing interstate highways. Further, the deterioration rates of pavements are selected from Fig. 2 based on the traffic levels in this work, without loss of generality, as mentioned in Section 5.1.

7.2. Network level risk

In the context of reliability-based decision problems, the risk is usually defined as an expected cumulative discounted failure state cost over the life-cycle, as in Andriotis and Papakonstantinou (2021). Using the same definition, we will only associate risk with bridge failures for this network, and the associated cost can be calculated as:

$$c_{\mathcal{R}\pi} = \mathbb{E}_{\theta_{1:T}} \left[\sum_{t=0}^T \gamma^t \left(c_{F_1} P_{F_{t+1}|a_{0:t}, \theta_{0:t}} + c_{F_2} \left(P_{F_{t+1}|a_{0:t}, \theta_{0:t}} - P_{F_t|a_{0:t}, \theta_{0:t}} \right) \right) \right] \quad (28)$$

where $c_{\mathcal{R}\pi}$ is the risk cost which has two components: (1) accruable cost (c_{F_1}) and (2) instantaneous cost (c_{F_2}). Instantaneous failure risk cost is the cost incurred upon the system entering a damage state and is not collected for the duration of the system's sojourn time in this damage state. It is herein taken as ten times the rebuilding cost of the bridge. The accruable cost is taken as twice the rebuilding cost of the bridge and is imposed at every time step during the relevant sojourn time. Two types of risks are considered: (i) risk of individual bridge failures (for

all network bridges) and (ii) system-level risk. The risk associated with bridge failures can straightforwardly be calculated based on their failure probability and the two costs mentioned above. This risk can be directly added to the cost function or objective function.

For the system risk, first, we define the system level risk and its failure modes 1–3. The loss of connectivity due to bridge failures in different parts of the network is considered a system failure. Mode 1 failure is considered when the network loses the connection between Gloucester County and Hampton connected via a single bridge. Mode 2 failure is defined as the failure of all three bridges connecting Hampton to the southern part of the network. Finally, mode 3 failure is defined as the simultaneous occurrence of mode 1 and mode 2 failures. These failure modes are shown in Fig. 7. Failed links and bridges are indicated in black, and the dotted line indicates the network partitioning due to disconnection.

It is critical to note that maintenance actions do not result in link closures; a link closure only occurs when a bridge between two nodes fails. Based on the failure probability of bridges, the failure probability of the three modes can be obtained in closed form using a combination of closed links. The two costs associated with each mode can be calculated as the sum of the failure costs of the involved bridges. The estimation of failure costs using instantaneous and accrued costs can be seen in (Andriotis & Papakonstantinou, 2021).

7.3. Delay costs

For transportation networks, maintenance and rehabilitation activities have associated user delay costs due to work zone-related restrictions (like lane closures, speed reductions, and others), which result in increased travel time. FHWA defines the user delay cost as the additional costs borne by commuters and the community at large due to work zone activity. User delay cost is an important ingredient of the life-cycle cost analysis of pavements and bridges, which also incorporates agency costs such as inspection and maintenance costs and other user-related costs. This study estimates user delay cost by first calculating the additional user time to commute due to the work zones. The model developed in (Vadakpat et al., 2000) is used for estimating the additional travel time during the maintenance period as:

$$T_d = \exp \left(1.67 + 0.0129(Trucks) + 0.000003(X - X')^2 + 0.00395(X - X') + 0.602(Lane) \right) \quad (29)$$

$$X' = (X + 48,600)/5$$

where T_d is the additional travel time (in vehicle-hours) incurred by all vehicles as the result of the work zone, $Trucks$ is a truck percentage in

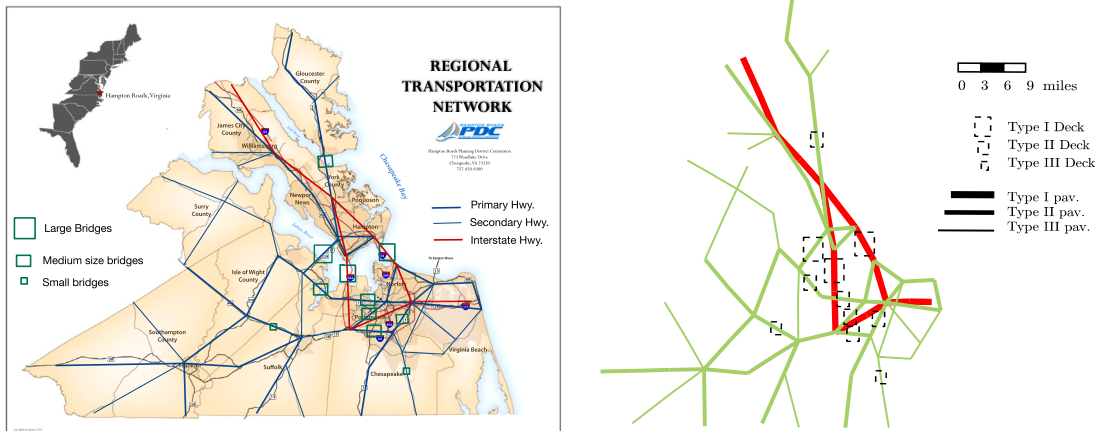


Fig. 6. (Left) Modified Hampton Roads network map Hampton Roads Planning District Commission (HRPDC) (2026). (Right) Network topology modeling of Hampton Roads: Green links are primary and secondary pavements, and red links are interstate highways.

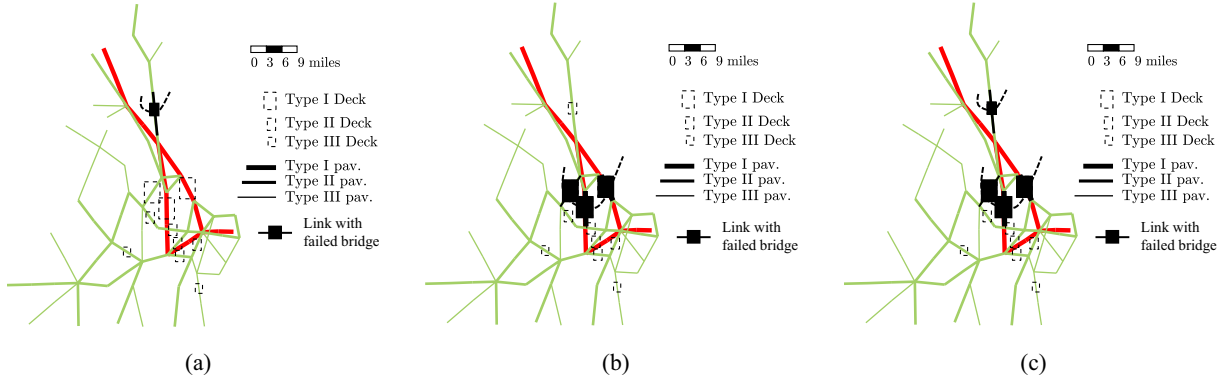


Fig. 7. Different configurations of system failures: (a) Mode 1, (b) Mode 2, (c) Mode 3.

Table 9

Average Daily Vehicle miles Traveled (DVMT) of the modeled Hampton Roads network [Daily Vehicle Miles Traveled \(DVMT\) \(2020\)](#).

Pavements	Total miles	DVMT mileage in 2019 (million miles)	Truck traffic (%)
Interstate (type I)	68.9	6.040	5.52
Primary (type II)	267.67	7.560	10.96
Secondary (type III)	145	0.486	1.85

the vehicle stream, X is the volume for which the delay is computed (as vehicles per hour), and $Lane$ is an indicator of which lane is blocked (1 if the right lane is closed, 0 if the left lane is closed).

To calculate the delay cost for the network using this model requires traffic distribution and USD costs/hr values for different vehicles. The traffic distribution is calculated based on the average Daily Vehicle Miles Traveled (DVMT) for the roads in Hampton Roads, as provided in [Daily Vehicle Miles Traveled \(DVMT\) \(2020\)](#) for the year 2019. The data for 2020 onwards is not used due to the pandemic-induced traffic disruption. Table 9 summarizes the traffic distribution for interstate, primary, and secondary pavements, however, explicit bridge data are not present in [Daily Vehicle Miles Traveled \(DVMT\) \(2020\)](#). We thus extrapolated traffic based on the total DVMT and the number of bridge miles present in the network. The truck percentages are assigned to type I-III bridges, the same as for interstate-secondary ones. Further, to use Eq. (29), the daily vehicular traffic volume is converted to hourly volume per lane-mile for the individual components. A pair of lanes is considered at a time to calculate the additional delay time, i.e., (a) $T_{d,R}$ when the right lane is closed, (b) $T_{d,L}$ when the left lane is closed. The following equation translates the obtained delay time to USD cost as:

$$c_D = (T_{d,L} + T_{d,R}) T_m (c_{car}(1 - Trucks) + c_{truck} Trucks) n_p L \quad (30)$$

where c_D is the delay cost in USD, T_m is the maintenance duration in hours, L is the component length in miles, n_p is the number of lane pairs, for example a 4-lane component would have two-lane pairs, $c_{car} = \$21.89$ is an average hourly wage for passenger cars, and $c_{truck} = \$29.65$ is an average hourly wage for trucks. These hourly wage rates depend upon the consumer price index and other macroeconomic indicators for a particular year; thus, the values above are based on the 2019 data reported by FHWA [Federal Highway Administration \(2020\)](#).

7.4. Network constraints

In the context of transportation asset management, performance measurement is the process of estimating progress toward preset goals or objectives using evidence such as aggregate measurements of conditions, smoothness of surfaces, and safety, as well as public perception

(e.g., public surveys). To offer the best possible transportation service within the constraints of available resources, federal and state agencies have established goals for transportation asset performance with a clear definition of performance measures. For instance, under the FHWA objectives for bridges, 23 CFR (Code of Federal Regulations) 490.411(a) [Federal Highway Administration and United States Department of Transportation \(2026a\)](#) states that no more than 10% of the National Highway System's (NHS) bridges by total deck area should be in poor condition (i.e., structurally deficient with a rating of ≤ 4). Similarly, 23 CFR part 490.315(a) subpart (c) [Federal Highway Administration and United States Department of Transportation \(2026b\)](#) mandates that no more than 5% of the state's NHS interstate lane miles be in poor condition. Additionally, the FHWA requires biannual inspections of bridges to maintain up-to-date bridge condition states, however, in this analysis, we are not adding this restriction, to solve the more challenging problem of simultaneously obtaining optimal inspection and maintenance policies.

Similarly, state DOTs impose performance targets on pavements in addition to the FHWA ones, such as the VDOT goal of no more than 18% of interstate and primary highways and 35% of secondary roadways being designated as deficient (i.e., CCI < 60). In terms of serviceability, no more than 15% of interstate and primary highways should be classified as poor in ride quality (i.e., IRI > 2.2 m/km). VDOT also aims to achieve a CCI of no less than 35 for the interstate system, [Virginia Department of Transportation \(2018, 2019\)](#). Apart from performance targets, there are also resource limits, such as the available budget, which cannot be exceeded. Hampton Roads districts have been granted a five-year budget of ~ 1.3 billion for 2021–2026 [Hampton Roads Transportation Planning Organization \(HRTPO\) \(2024\)](#). Summarizing, we are considering six performance constraints and one budget constraint, as follows:

1. For NHS bridges, poor condition decks by area should be less than 10% of the total deck area.
2. For NHS pavements, no more than 5% of the interstate lane-miles should be classified as deficient (i.e., CCI < 60 and IRI > 2.20 m/km).
3. No more than 18% of interstate and primary roadway pavement conditions should be classified as deficient (i.e., CCI < 60).
4. No more than 15% of interstate and primary roadways should be classified as deficient in terms of ride quality (i.e., IRI > 2.2 m/km).
5. No more than 35% of secondary roadways pavement condition should be classified as deficient (i.e., CCI < 60).
6. No more than 2% of interstate lane miles should have a CCI less than 37.
7. Hampton roads available budget for FY2021-2026 is \$ 1.3 billion, renewed every five years, discounted for every five year-cycle.

All six performance-related targets imposed by FHWA and VDOT are integrated as soft constraints in our DRL framework, for the studied Hampton Roads network, i.e., these performance goals will be met in an average sense, during the life of the considered network system, as mentioned in previous sections. Since it is infeasible to stochastically satisfy the VDOT performance goal of no pavement with CCI less than 35, this is changed from 0 to 2% in this work. Lastly, the available budget is considered a hard constraint and implemented as an augmented, fully observable state of the network. Additionally, the budget limit for each cycle is dynamically adjusted to account for the discounting of future values.

7.5. Condition-based maintenance and VDOT policy baselines

A Condition-Based Maintenance (CBM) policy is designed and optimized for performance comparisons in order to serve as a baseline against the obtained DRL solution. The most cost-effective policy is sought, to find the relevant thresholds that activate inspection and maintenance actions based on the condition of each component type, i.e., bridge, interstate, primary, and secondary pavements. The derived optimized policy involves the full suite of actions at every even time step, however, at every odd year, actions do-nothing and high-fidelity-inspection are deemed optimal. The CBM rules are given in Algorithm 2 and 3 for intact and non-intact starting condition states, respectively, as explained subsequently. These condition-based rules are localized with respect to the individual components, and cannot achieve a global optimum for the entire network. However, these rules are chosen such that they can satisfy all the system constraints imposed in the previous section as tightly as possible.

Another baseline largely based on the VDOT policy Chowdhury (2016) is also added for comparison. The VDOT policy assesses 100% of the interstate and primary systems and 20% of the secondary systems each year. The policy is divided into two stages. In the first stage, after the assessment of a component, depending on the current distress decision matrix and CCI filter, an action is recommended Chowdhury (2016). In this study, due to the unavailability of a distress decision matrix, this first stage recommendation is solely based on the CCI filter, as reported in Section 5.1.2. In the second stage, the recommended maintenance action may be updated based on the effective pavement age, Average Annual Daily Truck Traffic (AADTT), and Falling Weight Deflectometer (FWD). Different highway types have different decision trees with different sets of criteria and trigger values. For example, interstate highways utilize all three criteria (age, AADTT, and FWD, however due to the lack of data on FWD we are only using age and AADTT in this work), primary highways consider age and AADTT in their decision trees, and secondary highways decision trees are only based on AADT values. A second stage decision tree is shown in Fig. 8 as an example case, with DN as a first-stage recommended action for primary highways Chowdhury (2016). Fig. 8 shows the decision tree with starting node DN splitting into three different age groups, and the final decision is taken based on the different AADTT levels. It may be worth reminding that our action categories *Do-Nothing*, *Minor*, *Major*, *Reconstruction* correspond to the VDOT defined actions DN, PM, CM/RM, and RC, respectively, as mentioned in Section 5.1.4. Due to the introduced risk metrics for bridges in this work, we have again designed and optimized an annual CBM policy for bridge decks, to be used now in combination with the VDOT based rules for pavements. The derived most cost-effective policy rules in this case are given in Table A.1 in the Appendix, for both intact and non-intact starting condition states. Due to budget depletion in this VDOT-based policy case, for the non-intact starting conditions, a prioritization scheme is also applied in this case, at applicable time steps, as also described in the Appendix Table A.1.

In this study, DDMAC-CTDE is trained from scratch to evaluate the ability of the proposed framework to independently learn constrained inspection and maintenance policies. However, other policies, such as the ones mentioned here, could perhaps be used in the hope of

warm-starting the DRL training; for example by initializing the replay buffer with relevant generated trajectories or by using imitation-based pretraining. Yet, such strategies can often lead to non-desired performance since, for example, the limited threshold structure of a CBM policy may bias the DRL policy toward suboptimal behavior and limit its ability to exploit centralized training signals for coordinated system-level decisions. Along these lines, warm-starting the actual DRL problem training with relaxed constrained solutions, either through their neural network parameters and/or relevant trajectories, can often prove much more beneficial.

8. Results

This study considers a network from Hampton Roads, Virginia, with 96 components having a total number of $\sim 2 \times 10^{248}$ possible system states at any given time instant. Ten actions per component are considered with action labels 0-9, where {0,1,2} represent *Do Nothing*, *Minor*, and *Major Repair* with no-inspection, respectively, {3,4,5} action codes are the same maintenance actions with low-fidelity inspection, and action codes {6,7,8} represent the three respective maintenance actions combined with high-fidelity inspection. Finally, action code {9} indicates *Reconstruction*. The total number of available actions is 10^{96} for the entire network at each step. Out of 96 components, the network contains 85 pavement sections, and 11 important bridges of different sizes, as shown in Table B.2. The pavements are discretized with lengths ranging from 2.7 to 8.9 miles and a constant lane-width of 12 feet, with 12 interstate highways, 47 primary highways, and 26 secondary highway components. At $t = 0$, the network starts from (a) an intact state and (b) a non-intact condition state (as of 2021; details can be found in Appendix B), with an episode length of 20 years, and discount factors $\gamma = \gamma_h = 0.97$ and $\gamma_s = 1$. The two corresponding network cases are shown in Fig. 9. These networks are trained using the budget and the six performance constraints mentioned in Section 7.4. A terminal cost, dependent on the component's state, is added to the total cost to incentivize the agent to take maintenance actions even at the end of the policy horizon, adhering to engineering safety/serviceability requirements and assisting in avoiding constraint violations at the end of the considered horizon. The method of calculating this terminal cost is included in Appendix D. The training of the network is performed based on the details provided in the following subsection.

8.1. Experimental details

For the purposes of this numerical investigation here, the actor and critic networks consist of fully connected layers. Each bridge actor receives its local deck-condition belief vector and component deterioration rate, together with the shared budget or constraint state and the current time step. For pavement components, a fully observable binary reconstruction indicator denotes whether a long-duration replacement is still in progress, during which the component is temporarily unavailable for another intervention. Accordingly, each pavement actor receives its local IRI and CCI belief vectors, component deterioration rate, local reconstruction indicator, shared budget or constraint state, and current time step. The centralized critic receives the concatenated IRI, CCI, and deck-condition belief vectors of all pavement and bridge components, the deterioration rates of all components, the full pavement reconstruction-status vector, the system-level budget or constraint state, and the current time step. Each actor network (one per component) has two hidden layers with 200 neurons each, utilizing ReLU activation, and an output layer with 10 neurons using a *SoftMax* activation function. The centralized critic network features two hidden layers with 500 neurons each, also employing ReLU activation, and an output layer with 1 neuron using a *Linear* activation function. The training involves 11×10^5 episodes, each lasting 20 time steps, with a buffer memory size of 15,000 episodes. Exploration noise varies from 1.00 to 0.05 over 25,000 episodes, stabilizing at 0.02 after 100,000 episodes. Mini-batches of size 32 time steps are used, and network parameters are updated after each

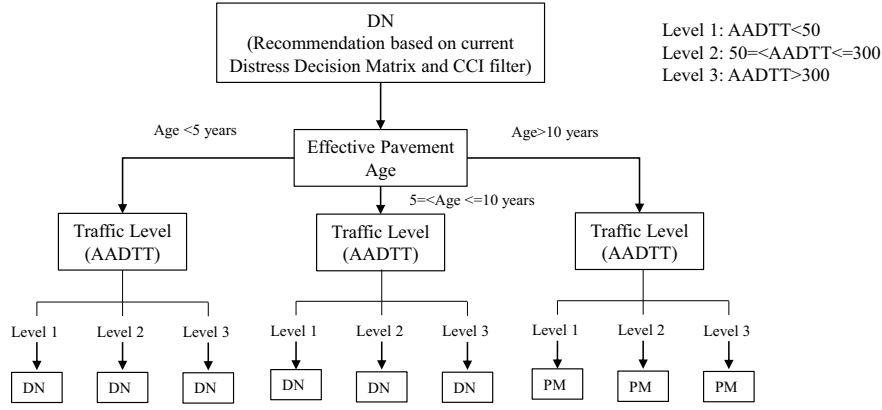


Fig. 8. Augmented decision tree for primary highways with Do Nothing (DN) as a stage one recommended action, along with trigger values based on effective age and Annual Average Daily Truck Traffic (AADTT).

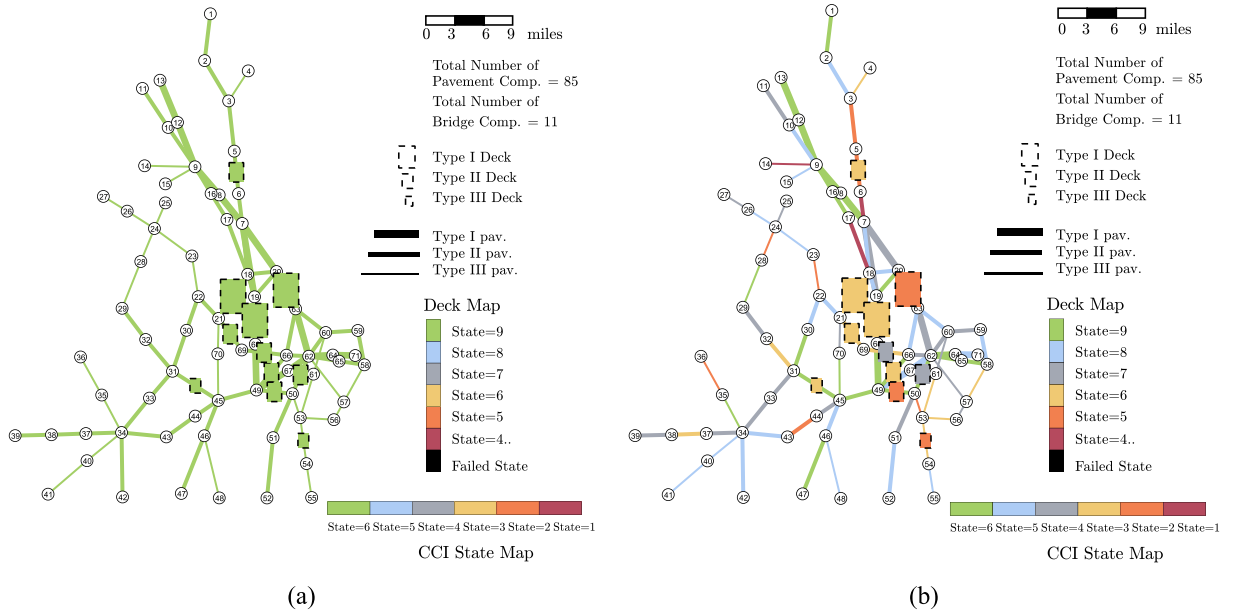


Fig. 9. Discretized network with 96 components consisting of 11 bridges and 85 highway sections starting with initial conditions corresponding to (a) intact state and (b) non-intact actual state as of 2021.

episode. The Adam optimizer is applied to both networks, with actor learning rate 0.0001, critic learning rate 0.001, and reducing the learning rate by half every 100,000 episodes until it stabilizes at 500,000 episodes. The loss function for the actor networks is the categorical cross-entropy, while the critic network uses a mean squared error function. The training is performed on a CPU with 96 cores and 364 GB of RAM.

8.2. DDMAC-CTDE and baseline solutions

Training for the DDMAC-CTDE network for 11×10^5 episodes took 8–9 days with our available computational resources, while the computed optimized baselines took 3–4 days for different instances. Fig. 10(a) & Fig. 10(b) show a few DDMAC-CTDE training curves (solid lines) for this network problem, starting with intact and non-intact states, respectively, along with the CBM policy (orange dashed dot line) discussed in Section A in the Appendix, and the VDOT policy (yellow dashed line) as explained before and largely specified in Chowdhury

(2016). As seen in these figures, the DDMAC-CTDE solution surpasses the baselines in both cases with different starting states.

The original DDMAC approach could not be evaluated in this benchmark setting, as it could not converge reliably due to the dimensionality of the problem. The main issue is the actor-side input dimension. Under the implemented state encoding, the Hampton Roads benchmark contains 85 pavement components and 11 bridge components. Each pavement component is represented by 6 CCI belief states and 5 IRI belief states, while each bridge component is represented by 9 deck-condition belief states. Thus, the relevant system belief consists of $85 \times (6 + 5) + 11 \times 9 = 1,034$ belief entries. In contrast, DDMAC-CTDE provides each actor only with its related local belief: a pavement actor receives $6 + 5 = 11$ inputs, while a bridge actor receives 9 inputs. With the actor architecture used here, consisting of two hidden layers with 200 neurons each, this reduces the first-layer actor weights from $1,034 \times 200 = 206,800$ in the original DDMAC information structure to $11 \times 200 = 2,200$ for pavement actors and $9 \times 200 = 1,800$ for

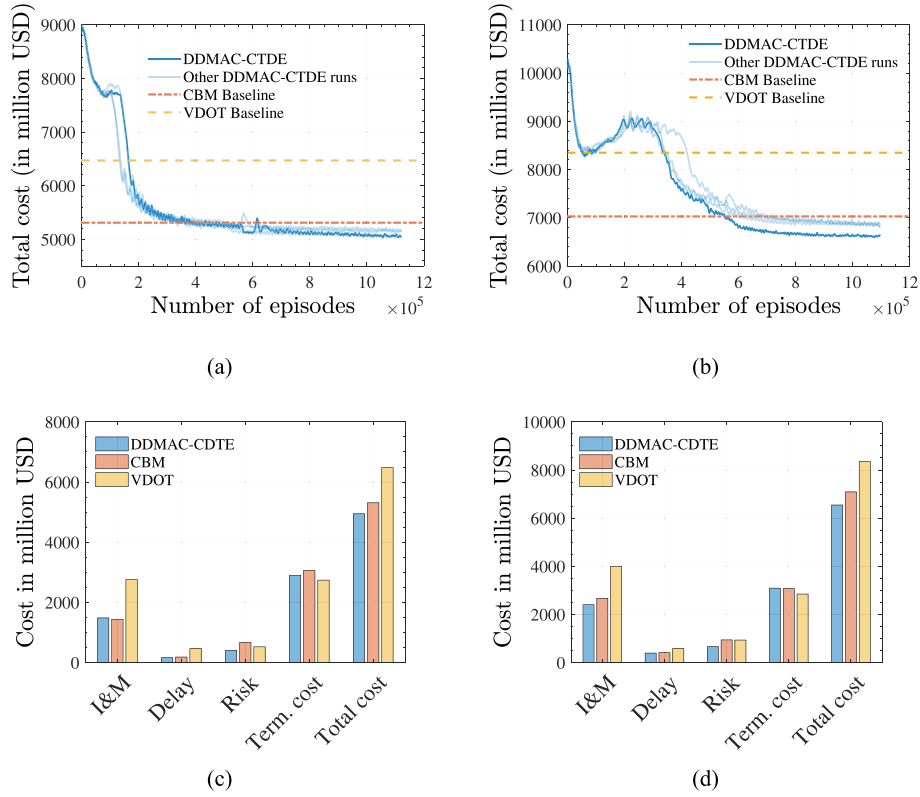


Fig. 10. Total life cycle costs (excluding terminal cost) comparison of the DDMAC-CTDE solution with CBM and VDOT policy baselines (top) for the network with (a) intact starting state; (b) non-intact starting state. Comparison of the total cost and its constituents with CBM and VDOT policy baselines (bottom) for the network with (c) intact starting state; (d) non-intact starting state.

Table 10

Comparison of objectives and imposed constraints using different policies for intact and non-intact starting condition states, where I-, P-, and S-Hwy are interstate, primary and secondary highways.

Objectives & constraints	Intact starting state			Non-intact starting state		
	DDMAC-CTDE	CBM policy	VDOT policy	DDMAC-CTDE	CBM policy	VDOT policy
Total cost incurred (billion USD)	4.94 ($\pm 1.5 \times 10^{-3}$)	5.31(+7.5%) ($\pm 3.4 \times 10^{-3}$)	6.47(+31%) ($\pm 4.7 \times 10^{-3}$)	6.55 ($\pm 2.9 \times 10^{-3}$)	7.03(+7.3%) ($\pm 4.4 \times 10^{-3}$)	8.35(+27%) ($\pm 7.6 \times 10^{-3}$)
Total cost incurred w/o terminal cost (billion USD)	2.04 ($\pm 1.2 \times 10^{-3}$)	2.25(+10%) ($\pm 3.4 \times 10^{-3}$)	3.74(+83%) ($\pm 4.5 \times 10^{-3}$)	3.45 ($\pm 2.7 \times 10^{-3}$)	3.97(+15%) ($\pm 4.3 \times 10^{-3}$)	5.50(+59%) ($\pm 7.4 \times 10^{-3}$)
CCI < 60 and IRI > 2.2 m/km for I-Hwy (% , cap 5%)	2.94(± 0.03)	2.82(± 0.04)	0.00(± 0.00)	1.34(± 0.02)	1.61(± 0.02)	3.39(± 0.07)
CCI < 35 of I-Hwy (% , cap 2%)	2.68(± 0.02)	0.70(± 0.01)	0.02(± 0.00)	1.50(± 0.02)	0.68(± 0.01)	3.80(± 0.07)
CCI < 60 for I and P-Hwy (% , cap 18%)	5.27(± 0.02)	4.41(± 0.02)	0.05(± 0.00)	17.65(± 0.04)	12.67(± 0.02)	5.80(± 0.04)
IRI > 2.2 m/km for I and P-Hwy (% , cap 15%)	14.76(± 0.04)	12.88(± 0.05)	11.23(± 0.04)	15.40(± 0.04)	11.49(± 0.04)	15.65(± 0.07)
CCI < 60 for S-Hwy (% , cap 35%)	10.13(± 0.04)	7.79(± 0.03)	0.72(± 0.01)	33.00(± 0.08)	28.18(± 0.04)	37.86(± 0.13)
Decks with condition rating ≤ 4 (% , cap 10%)	0.95(± 0.02)	4.76(± 0.05)	2.22(± 0.04)	8.33(± 0.05)	8.79(± 0.06)	15.85(± 0.15)

Note: The estimated 95% confidence bounds, assuming Gaussian distributions, are provided in parentheses, calculated based on 10^4 simulations. The bold values indicate the difference in costs for CBM and VDOT with respect to the DDMAC-CTDE solution.

bridge actors. Across all 96 actors, this corresponds to an approximately 80-fold reduction in first-layer actor weights. Attempts to train the original DDMAC information structure thus led to unstable learning and non-convergent policies within the available training budget, whereas DDMAC-CTDE trained much more reliably while retaining centralized value guidance through the critic. This should be interpreted as a practical scalability and training-stability motivation, while recognizing that local actor inputs may introduce a decentralization trade-off for some coupled and strongly coupled systems Bhustali et al. (2026).

Fig. 10(c) & Fig. 10(d) provide the histogram comparison of converged costs for the three utilized policies. The various costs such as inspection and maintenance, user delay, life-cycle risk, and terminal cost are distinct constituents of the total cost. These plots indicate that the DDMAC-CTDE achieves the lowest total cost metrics. The summarized results and comparisons are further presented in Table 10. The reported results represent average values derived from 10^4 simulated policies using the trained networks and optimized baselines. The solid dark blue instance in Fig. 10(a) and Fig. 10(b) denotes the trained

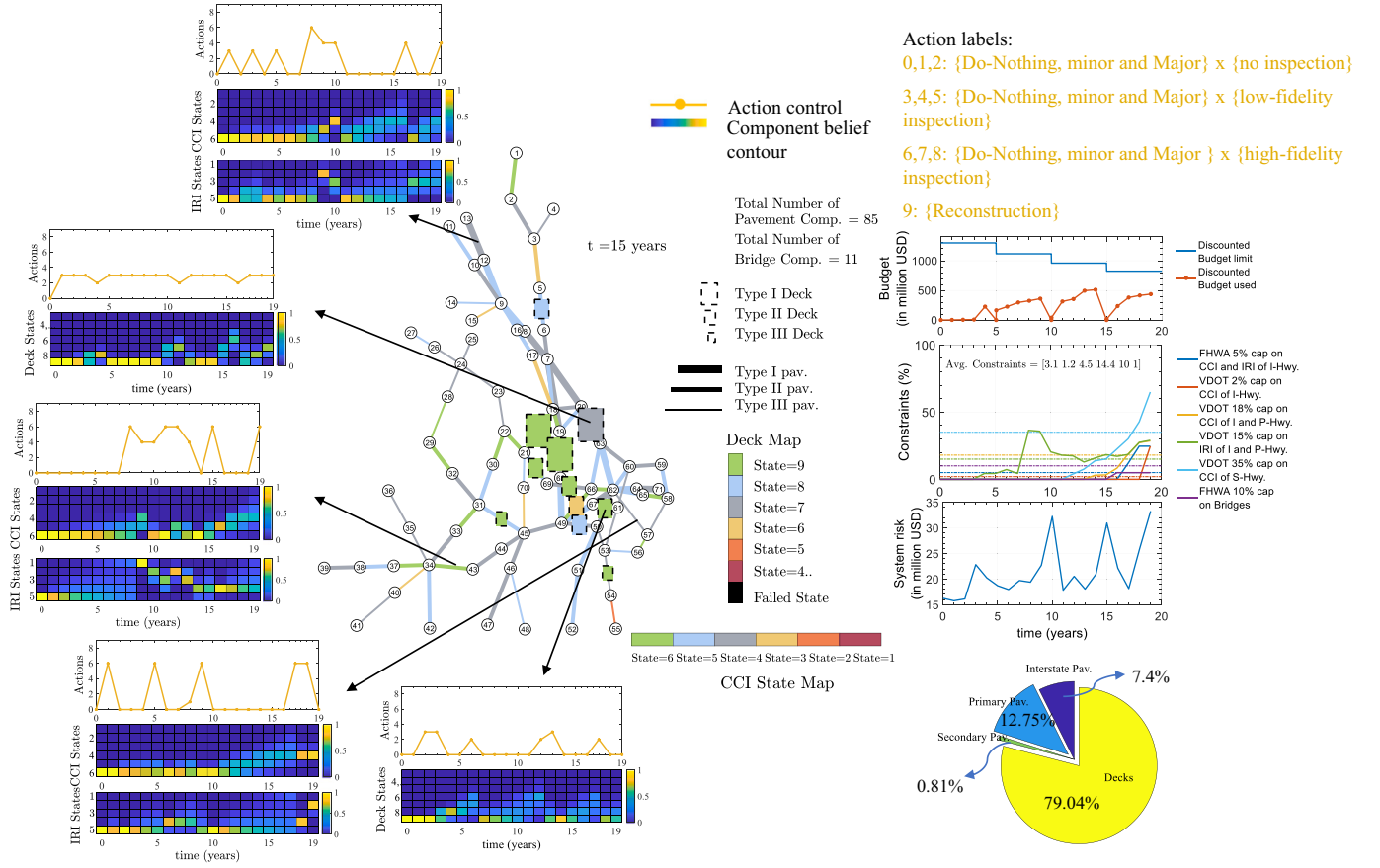


Fig. 11. Life-cycle realization of the learned DDMAC-CTDE policy for the Hampton Roads network, starting from an intact state.

network employed for the mentioned simulations. Table 10 outlines the objectives and six performance constraints in comparison with CBM and VDOT policies. DDMAC-CTDE demonstrates a significant advantage, with an estimated total life-cycle cost of 4.94 billion USD for the intact starting state, compared to 5.31 billion USD for the CBM policy (7.5% more expensive) and 6.47 billion USD for the VDOT policy (31% more expensive). In the case of the non-intact starting state, DDMAC-CTDE incurs a cost of 6.55 billion USD, while CBM and VDOT policies result in 7.3% and 27% higher life-cycle costs, respectively, compared to DDMAC-CTDE. Excluding the terminal cost, the proposed method achieves even more compelling cost savings. Specifically, for the intact case, CBM and VDOT policies are approximately 10% and 83% more expensive, respectively, compared to DDMAC-CTDE. Similarly, for the non-intact case, CBM and VDOT policies are approximately 15% and 59% more expensive, respectively, than the DDMAC-CTDE policy. Additionally, DDMAC-CTDE results in terminal costs comparable to CBM- and VDOT-policies, without adhering to a conservative maintenance approach. Furthermore, it successfully meets the posed targets and constraints, overall.

Table 10 also compares the average performance over 10^4 simulations in terms of the six different constraints detailed in Section 7.4. The performance constraints are in the rows of the table, and I, P, and S Hwy are the abbreviations for interstate, primary, and secondary highways, respectively. As noted in the table, the DDMAC-CTDE solutions satisfy almost all the constraints for both starting states, except for the soft constraints $CCI < 35$ of I-Hwy (2.68% surpassing the 2% target), for the intact state case, and $IRI > 2.2$ m/km for I and P-Hwy (15.4% exceeding the 15% target) for the non-intact state case. In comparison, the CBM

based policy satisfies all the constraints for both intact and non-intact cases, while the VDOT-recommended policy satisfies all the constraints for the intact initial state scenario. For the non-intact case, the VDOT policy closely meets the constraints or exceeds them slightly in certain cases related to pavements, particularly for the targets for $CCI < 35$ for I-Hwy (3.80% exceeding the 2% target), $IRI > 2.2$ m/km for I and P-Hwy (15.65% exceeding the 15% target) and S-Hwy $CCI < 60$ (37.86% surpassing the 35% target). For the bridge decks target of condition ratings ≤ 4 , the implemented VDOT policy for the non-intact starting state significantly surpasses the 10% target by 5.85%, mainly as a result of the depletion of the budget, which was only observed in this case among all utilized policies and starting conditions.

To better understand how DDMAC-CTDE policies change over time, detailed policy realizations for some representative components are given in Figs. 11 and 12 for intact and non-intact starting states. The figures illustrate actions generated by one of the instances of the optimal policy and the evolution of belief states as a contour. Figs. 11 and 12 also depict the discounted budget utilization over time, as well as the discounted 5-year budget for each cycle. The budget is a hard constraint that the agents are not allowed to exceed, and this requirement is satisfied by the resulting solutions. Furthermore, it can be noticed that the agents do not utilize the full budget in both cases, since they can optimally satisfy the performance constraints with minimum life-cycle cost. In comparison, the VDOT policy for the non-intact starting state exhausted the entire budget and did not satisfy various performance constraints. Additionally, the temporal evolution of six performance constraints is presented in Figs. 11 and 12, in relation to pavement and bridge condition states. It is seen that the six performance thresholds are

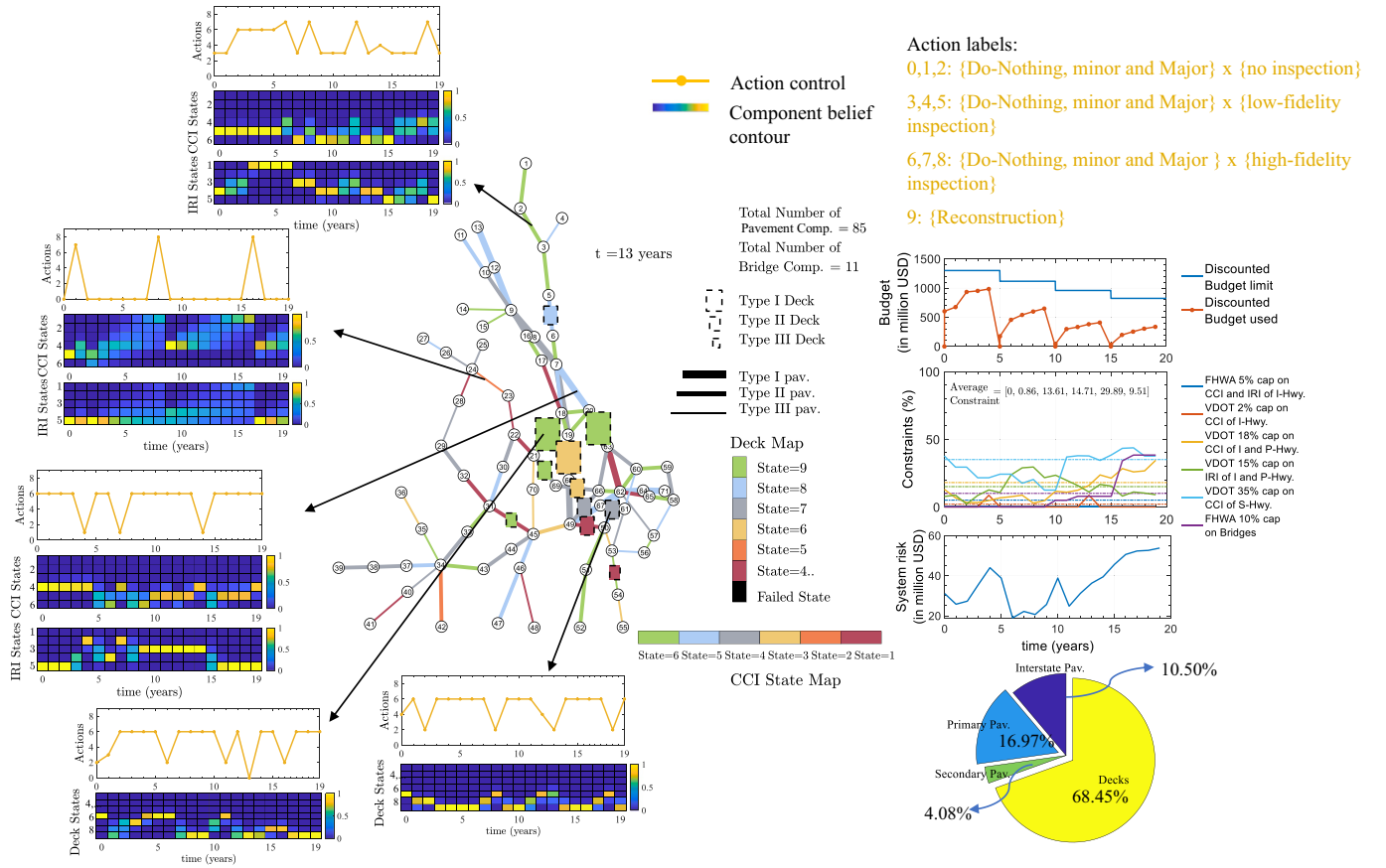


Fig. 12. Life-cycle realization of the learned DDMAC-CTDE policy for the Hampton Roads network, starting from the actual state as of 2021.

very well respected for the most part, except at the end of the life cycle. However, for these indicative realizations, the performance constraints are never violated on average over the life-cycle. The time variation of total risk cost, associated with individual bridges and the three modes of system risk, is shown as well. Finally, a pie chart illustrating the cost distribution for that realization among various types of pavements and bridges is also presented.

Plots with control actions represent the actions taken over time. The maintenance actions, taken at each time step, update the current belief of the system, as manifested in the next time step. The evolution of contour plots in the case of pavements shows current beliefs for both CCI and IRI states, and the current belief states at each step for two bridge decks are also shown. For example, in Fig. 11 the agent takes action 1 at $t = 8$ years for a secondary pavement, and the belief gets updated according to the *Minor Repair* action and the environmental effect transitions. Accordingly in Fig. 12, the agent is shown to take action 8 at $t = 8$ years for a secondary pavement, and then the updated beliefs of CCI and IRI are shown at $t = 9$ years, incorporating inspection outcomes, maintenance actions, and environmental transition effects.

As seen in Figs. 11 and 12, control actions are compatible with belief states. For example, in Fig. 11 the agents initially choose *Do-Nothing* actions since the belief states for both pavements and bridges initiate in the intact condition. As the conditions gradually deteriorate, more interventions are considered. However, this is not the case with the non-intact starting state example, where interventions need to start from the beginning of the planning horizon. Further, the inspection and maintenance actions at the end of the horizon are justified in both cases due to the state-associated terminal cost and the present risk for bridges.

It is also observed through the policy simulations that the agents maintain the type I bridges more systematically and consistently choose inspection actions for them to have better informed state estimates. This is because type I bridges have their individual failure risk, as well as mode 1 and mode 3 system failure risks associated with them. Similarly, importance is given to more essential pavement components. For example, in Fig. 12, the maintenance and inspection actions are more frequent for interstate than primary components, as well as for primary pavements with respect to secondary ones.

From the pie charts, shown in Figs. 11 and 12, it is observed that cost distributions are heavily skewed (as much as 79% for intact start and 68.45% for non-intact start) toward the bridge components, due to their high maintenance costs, associated risk costs, and lower traffic delay costs. The cost is even more skewed for the intact versus non-intact start case due to the relatively slower initial deterioration rate of pavements in comparison to bridges in this work. Thus, the pavements with intact starting state conditions do not need significant maintenance activity until the later stages of their life-cycle. Among pavements, primary highways absorb the largest part of the budget, as they outnumber the other components in the network (47 in total). Figs. 11 and 12 also show the evolution of the system risk with time. In both cases, the risk is minimal initially and increases with time, with downward jumps mainly due to the maintenance actions taken for bridges, especially of type I.

9. Conclusions

This study develops a new stochastic optimal control deep reinforcement learning (DRL) paradigm for inspecting and maintaining

deteriorating transportation network systems under partial information and constraints. It establishes a comprehensive modeling framework for pavements and bridges using Partially Observable Markov Decision Process (POMDP) principles, in the most complex and detailed application of DRL to this type of system optimization problem, to the best of our knowledge. A large transportation network with 96 components situated in the state of Virginia, USA, is studied, and due to an exhaustively intractable number of system state and action combinations, the problem is solved with a newly developed DRL algorithmic approach. In particular, a Deep Decentralized Multi-agent Actor-Critic with Centralized Training and Decentralized Execution (DDMAC-CTDE) is presented, which uses sparse parameterizations and local component state information for actor networks to obtain near-optimal solutions. Two indices, the Critical Condition Index (CCI) and the International Roughness Index (IRI), are used to determine the pavement structural condition and serviceability, and the associated transition probabilities for both indices are determined. For bridges, the deck performance is utilized since it deteriorates at a faster rate and has a higher maintenance cost. Many operational, budget, and performance-based constraints are also imposed, by means of hard and soft constraints in the DRL formulation. The framework is successfully demonstrated to offer significantly better solutions than heuristically optimized Condition Based Maintenance (CBM) and VDOT recommended policy baselines, which are 7.5% and 31% more expensive, respectively, starting from an intact condition state. The proposed framework also effectively handles all performance targets and constraints imposed by FHWA and VDOT. In addition to the intact starting condition state, the Hampton Roads network's 2021 conditions are also considered as starting states. In this case, DDMAC-CTDE policies again outperform the considered baselines with respect to the total cost, with the CBM and VDOT policies being 7.3% and 27% more expensive, respectively. Overall, this work importantly contributes toward DRL-based inspection and maintenance policies for transportation networks and related infrastructure systems.

Future work will further investigate the scalability of the proposed framework to regional and state-wide transportation networks, building on recent efforts on district-level bridge networks with thousands of assets Arcieri et al. (2025); Bhattacharya et al. (2025). Such extensions

may require more efficient simulation techniques, approximate system-level coordination, parameter sharing, graph-based and/or hierarchical architectures. Other future directions can include robust optimization approaches Arcieri et al. (2023, 2024), dynamic traffic effects, and incorporation of traffic control actions into the entire management framework.

CRedit authorship contribution statement

M. Saifullah: Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **K.G. Papakonstantinou:** Writing – original draft, Writing – review & editing, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **A. Bhattacharya:** Writing – review & editing, Software, Methodology, Investigation, Data curation. **S.M. Stoffels:** Writing – review & editing, Validation, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization. **C.P. Andriotis:** Writing – review & editing, Software, Methodology, Investigation, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: K.G. Papakonstantinou, M. Saifullah, and C. P. Andriotis have patent #US Patent 12,346,795 issued. Other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors acknowledge the support of the [U.S. National Science Foundation](#) under Grants No. [CMMI 1751941](#) and [2053620](#), and the Center for Integrated Asset Management for Multimodal Transportation Infrastructure Systems, U.S. DOT Region 3 University Center. Dr. Andriotis would further like to acknowledge the support of the TU Delft AI Labs program.

Appendix A. Optimal condition based maintenance (CBM) policy**Algorithm 2:** Optimal Condition Based Maintenance (CBM) policy for intact starting condition state.

```

initialize total cost metric  $\bar{c} = 0$ ;
while  $i < \text{max\_episodes}$  do
   $t = 0$ ;
  initialize component beliefs  $\mathbf{b}_t = [\mathbf{b}_t^{(i)}]_{i=1}^N$ , observations
   $\mathbf{o}_t = [\mathbf{o}_t^{(i)}]_{i=1}^N$ ;
  while  $t < T$  do
    if  $t$  is even then
      select action  $a_t = 6$  for all components;
    else
      if component = bridge then
        select action  $a_t = [0, 6, 8, 6, 8, 8, 9]$  if the observed
        states are  $[9, 8, 7, 6, 5, \{4, \dots\}, \text{failure}]$ ,
        respectively;
      else if component = interstate then
        select action  $a_t = [0, 6, 7, 6, 7, 8]$  if the observed CCI
        states are  $[6, 5, 4, 3, 2, 1]$ , respectively;
      else if component = primary then
        select action  $a_{t,1} = [0, 3, 3, 4, 4, 5]$  if the observed
        CCI states are  $[6, 5, 4, 3, 2, 1]$ ;
        select action  $a_{t,2} = [0, 3, 4, 4, 5]$  if the observed IRI
        states are  $[5, 4, 3, 2, 1]$ ;
        if  $a_{t,1} \geq a_{t,2}$  then
          select action  $a_t = a_{t,1}$ ;
        else
          select action  $a_t = a_{t,2}$ ;
        end
      else if component = secondary then
        select action  $a_t = [0, 0, 3, 4, 4, 4]$  if the observed CCI
        states are  $[6, 5, 4, 3, 2, 1]$ , respectively;
    end
    Estimate total cost  $\bar{c}_{b,t}$  based on the current belief  $\mathbf{b}_t$ 
    accounting for the budget and terminal cost  $\bar{c}_{s_T}$ ;
    if  $t \neq T$  then
       $\bar{c}_{b,t} = \bar{c}_b$ ;
    else
       $\bar{c}_{b,t} = \bar{c}_b + \bar{c}_{s_T}$ ;
    end
    Estimate next belief state  $\mathbf{b}_{t+1}$ , and collect the next
    observation  $\mathbf{o}_{t+1}$ ;
    Assign  $\mathbf{b}_t \leftarrow \mathbf{b}_{t+1}$ ,  $\mathbf{o}_t \leftarrow \mathbf{o}_{t+1}$ ;
  end
   $\bar{c} = \bar{c} + \sum_{t=0}^{T-1} \bar{c}_{b,t}$ ;
end
 $\bar{c} = \bar{c} / \text{max\_episodes}$ ;

```

Note: Actions 0,1,2 are Do Nothing, Minor Repair, Major Repair with no-inspection; 3, 4, 5 are Do Nothing, Minor Repair, Major Repair with low fidelity inspection; 6, 7, 8 are Do Nothing, Minor Repair, Major Repair with high fidelity inspection; and 9 is Reconstruction.

Algorithm 3: Optimal Condition Based Maintenance (CBM) policy for non-intact starting condition state.

```

initialize total cost metric  $\bar{c} = 0$ ;
while  $i < \text{max\_episodes}$  do
   $t = 0$ ;
  initialize component beliefs  $\mathbf{b}_t = [\mathbf{b}_t^{(i)}]_{i=1}^N$ , observations
   $\mathbf{o}_t = [\mathbf{o}_t^{(i)}]_{i=1}^N$ ;
  while  $t < T$  do
    if  $t$  is even then
      select action  $a_t = 6$  for all components;
    else
      if component = bridge then
        select action  $a_t = [0, 6, 8, 6, 8, 8, 9]$  if the observed
        states are  $[9, 8, 7, 6, 5, \{4, \dots\}, \text{failure}]$ ,
        respectively;
      else if component = interstate then
        select action  $a_t = [0, 6, 7, 8, 8, 8]$  if the observed CCI
        states are  $[6, 5, 4, 3, 2, 1]$ , respectively;
      else if component = primary then
        select action  $a_{t,1} = [0, 3, 3, 4, 4, 5]$  if the observed
        CCI states are  $[6, 5, 4, 3, 2, 1]$ ;
        select action  $a_{t,2} = [0, 3, 4, 4, 5]$  if the observed IRI
        states are  $[5, 4, 3, 2, 1]$ ;
        if  $a_{t,1} \geq a_{t,2}$  then
          select action  $a_t = a_{t,1}$ ;
        else
          select action  $a_t = a_{t,2}$ ;
        end
      else if component = secondary then
        select action  $a_t = [0, 0, 3, 4, 4, 5]$  if the observed CCI
        states are  $[6, 5, 4, 3, 2, 1]$ , respectively;
    end
    Estimate total cost  $\bar{c}_{b,t}$  based on the current belief  $\mathbf{b}_t$ 
    accounting for the budget and terminal cost  $\bar{c}_{s_T}$ ;
    if  $t \neq T$  then
       $\bar{c}_{b,t} = \bar{c}_b$ ;
    else
       $\bar{c}_{b,t} = \bar{c}_b + \bar{c}_{s_T}$ ;
    end
    Estimate next belief state  $\mathbf{b}_{t+1}$ , and collect the next
    observation  $\mathbf{o}_{t+1}$ ;
    Assign  $\mathbf{b}_t \leftarrow \mathbf{b}_{t+1}$ ,  $\mathbf{o}_t \leftarrow \mathbf{o}_{t+1}$ ;
  end
   $\bar{c} = \bar{c} + \sum_{t=0}^{T-1} \bar{c}_{b,t}$ ;
end
 $\bar{c} = \bar{c} / \text{max\_episodes}$ ;

```

Note: Actions 0,1,2 are Do Nothing, Minor Repair, Major Repair with no-inspection; 3, 4, 5 are Do Nothing, Minor Repair, Major Repair with low fidelity inspection; 6, 7, 8 are Do Nothing, Minor Repair, Major Repair with high fidelity inspection; and 9 is Reconstruction.

Table A.1

Considered bridge CBM policy in the VDOT-based baseline case.

Starting condition	Policy based on bridge conditions	Prioritization
Intact	Select action $a_i = [0, 6, 6, 8, 6, 8, 9]$ if the observed states are $[9, 8, 7, 6, 5, \{4, \dots\}, \text{failure}]$, respectively	Not required
Non-intact	Select action $a_i = [0, 6, 8, 8, 8, 8, 9]$ if the observed states are $[9, 8, 7, 6, 5, \{4, \dots\}, \text{failure}]$, respectively	Due to insufficient budget in the non-intact starting condition case, bridges are prioritized based on their associated individual risk cost. The first eight bridges with the highest risk costs are then selected at each time step for implementation of inspection and maintenance actions, contingent upon budget availability. Relevant actions for all other components of the network are then selected in a randomized priority, until the entire budget is utilized.

Appendix B. Condition of the Hampton roads network

The existing condition of the Hampton Roads network can be obtained from the VDOT dashboard and ArcGIS map [Virginia Department of Transportation \(2026a,2026b\)](#). For bridge decks, these conditions are reported as a general condition rating from 9 to 1, as shown in [Table B.2](#). For pavements, the CCI ratings are provided for interstate pavements, however, only the condition distribution on an aggregate level is provided for primary and secondary pavements, as shown in [Table B.3](#). Random sampling is thus performed in this work, using the provided distribution, to obtain the initial condition states for the primary and secondary highways.

For IRI states, VDOT only provides the mean IRI value and the percentage of deficient pavements for all types of highway systems (type I-III), as presented in [Table B.4](#). VDOT uses different definitions of deficiency for different pavement types. Here, we considered, for simplicity, pavements as deficient if $IRI > 2.20$ m/km or 140 inch/mile for all pavement types (I-III), which is consistent with the definition given by VDOT for interstate highways (i.e., if the pavement is in IRI state 1 or 2 it will be classified as deficient). For sampling, a lognormal distribution is chosen here, as IRI has continuous positive values. The fitted lognormal distribution and the obtained mean (μ), and standard deviation (σ) in the log space are reported in [Table B.4](#).

Along with the condition states, the component effective age is needed to acquire the deterioration rates for non-stationary indices. For the CCI, the age distribution is taken from VDOT ([Chowdhury, 2016](#)), which is shown in [Figs. B.1\(a\)](#) and [Fig. B.1\(b\)](#) for interstate and primary highways, respectively. For secondary pavements, the same distribution as for primary pavements is utilized, but with a three-year shift. This assumption is based on the fact that VDOT inspects 20% of secondary

Table B.3

CCI state distribution for primary and secondary pavements.

Pavements	V. poor	Poor	Fair	Good	V. good	Excellent
CCI distribution of primary pavements	0.091	0.057	0.135	0.241	0.215	0.261
CCI distribution of secondary pavements	0.148	0.102	0.109	0.272	0.193	0.176

Table B.4

Statistics of IRI values for different pavement types, and mean and standard deviation of the fitted lognormal distribution.

Pavement	Mean IRI (m/km)	Deficient pavement (%)	μ	σ
Interstate	1.264	4.30%	0.169	0.361
Primary	1.485	12.20%	0.3115	0.41
Secondary	2.196	35.00%	0.4933	0.766

highways annually, compared to the annual assessment of primary pavements, thus accordingly assuming that maintenance activities for the secondary system will be at least 2.5 years behind. Because age data for decks are unavailable, the same age distribution as for the interstate pavements is used for type-I bridges, while the age distribution of primary pavements is utilized for the remaining bridges. Finally, sampling is performed over the corresponding distributions to determine the component effective ages.

Table B.2

Bridge geometric information and condition state as of 2021.

S.no.	Name	Starting node	End node	Length (miles)	Number of lanes	Deck Condition Rating (GCR 1-9)	Sampled deck age (years)
1	James river (type I)	18	21	4.42	4	6	5
2	MMM bridge tunnel (type I)	19	68	4.63	4	6	7
3	Hampton roads bridge tunnel (type I)	20	63	3.50	4	5	5
4	Coleman memorial bridge (type II)	5	6	0.71	4	6	9
5	Bridge_RD (type II)	21	69	0.71	2	6	9
6	High rise bridge (type II)	49	50	0.94	4	5	12
7	Berkley bridge (type II)	49	66	0.40	8	6	10
8	Gilmerton bridge (type II)	50	62	0.42	4	7	1
9	W Norfolk bridge (type II)	66	69	0.63	4	7	3
10	Bridge on Nansemond River Suffolk (type III)	31	45	0.20	4	6	13
11	Great bridge bypass (type III)	53	54	0.56	4	5	7

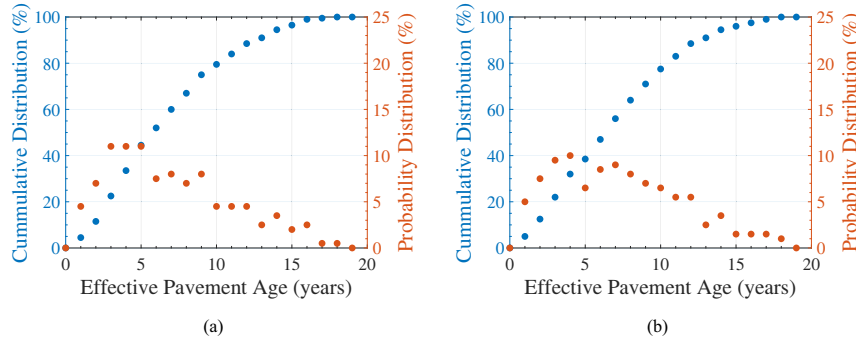


Fig. B.1. Age distribution of (a) interstate and (b) primary highways according to VDOT (Chowdhury, 2016).

Appendix C. Gamma process coefficients

The gamma process coefficients $f(t)$ and $g(t)$ from section 5.1.1 are presented in Table C.5 for $t = 1, 2, 3 \dots 20$ years.

Appendix D. Terminal costs

The terminal cost may be defined as the cost borne due to the condition state of the components at the end of the policy horizon, this is expressed through the value function at the last step $V^\pi(s_T)$, where s_T is the state at time step T , for finite-horizon problems of length T . If we consider no terminal costs/constraints (i.e., $V^\pi(s_T) = 0$), the agent usually refrains from performing any action at the end of the horizon, as there are no after-effects/benefits of these actions, except for the bridges in this work. However, in any engineering system problem, such as civil infrastructure maintenance problems, it is important to consider the aftermath of every decision-cycle. Thus, this terminal cost is added to incentivize the agents to still take inspection and maintenance actions at the end of the horizon. These actions also prevent violations of constraints, while seeking minimization of life-cycle costs and preservation of infrastructure beyond the current decision horizon.

For an arbitrary system terminal state, the terminal cost can be theoretically connected to an infinite horizon problem, for all components. However, due to computational infeasibility, the terminal cost is calculated here using a representative component from each type of pavement and bridge, and, for further simplification, the CBM policy in Algorithm 2, & 3 is chosen, with an unlimited budget, to approximate

the remaining value of the system components at the end of the decision horizon. The terminal cost/value function $V^{\pi_{CBM}}(s_T)$ is thus computed as:

$$c_{s_T} = \alpha V^{\pi_{CBM}}(s_T) = \alpha \mathbb{E}_{s \in S, a \sim \pi_{CBM}} \left[\sum_{t=T}^{\infty} \gamma^t c_t | s_T \right] \quad (D.1)$$

where π_{CBM} is a CBM action policy, c_t is the total cost that constitutes inspection and maintenance cost, user delay, and risk costs, and α is a hyperparameter determining the overall effect of the remaining value of the network. As this remaining cost naturally has a higher magnitude than other costs, the α value is carefully chosen to be 0.55 here, consistent with other utilized costs, and so that the computed policy is not influenced excessively, particularly at the end of the horizon. This cost is then directly added to the total cost at the last step of the life cycle, $t = T$.

Appendix E. Tables related to modeling of bridges and pavements

This section presents the rest of the matrices used for the pavement and deck models. Table E.1 provides the detailed maintenance actions for the pavements based on VDOT Chowdhury (2016). Tables E.2 and E.3 provide the observation probabilities for six CCI states, for low- and high-fidelity inspections, respectively, based on the model described in Section 5.1.3, while Table E.10 depicts their costs. Similarly, Table E.11 describes the costs and observation probabilities for IRI with 3 different levels of fidelity. Three maintenance actions (*Minor Repair*, *Major Repair*,

Table C.5
Variables $f(t)$ and $g(t)$ for gamma process for calculating CCI transition probabilities.

Time (years)	Traffic level (A)		Traffic level (B)		Traffic level (C)		Traffic level (D)		Traffic level (E)	
	$f(t)$	$g(t)$	$f(t)$	$g(t)$	$f(t)$	$g(t)$	$f(t)$	$g(t)$	$f(t)$	$g(t)$
1	0	0.178	0	0.280	0	0.349	0	0.420	0	0.473
2	0.922	0.101	1.217	0.158	1.394	0.197	1.560	0.237	1.679	0.267
3	0.649	0.061	0.857	0.096	0.981	0.119	1.098	0.143	1.182	0.161
4	0.475	0.038	0.628	0.059	0.718	0.074	0.804	0.089	0.866	0.100
5	0.354	1.135	0.467	1.350	0.535	1.469	0.599	1.577	0.644	1.651
6	12.759	0.952	12.759	1.133	12.759	1.233	12.759	1.323	12.759	1.386
7	12.759	0.801	12.759	0.953	12.759	1.037	12.759	1.113	12.759	1.166
8	12.759	0.675	12.759	0.804	12.759	0.874	12.759	0.938	12.759	0.982
9	12.759	0.570	12.759	0.678	12.759	0.738	12.759	0.792	12.759	0.829
10	12.759	0.481	12.759	0.573	12.759	0.623	12.759	0.669	12.759	0.700
11	12.759	0.407	12.759	0.484	12.759	0.527	12.759	0.566	12.759	0.592
12	12.759	0.344	12.759	0.410	12.759	0.446	12.759	0.479	12.759	0.501
13	12.759	0.292	12.759	0.347	12.759	0.378	12.759	0.405	12.759	0.424
14	12.759	0.247	12.759	0.294	12.759	0.320	12.759	0.343	12.759	0.359
15	12.759	0.209	12.759	0.249	12.759	0.271	12.759	0.291	12.759	0.305
16	12.759	0.177	12.759	0.211	12.759	0.230	12.759	0.247	12.759	0.258
17	12.759	0.151	12.759	0.179	12.759	0.195	12.759	0.209	12.759	0.219
18	12.759	0.128	12.759	0.152	12.759	0.165	12.759	0.177	12.759	0.186
19	12.759	0.108	12.759	0.129	12.759	0.140	12.759	0.151	12.759	0.158
20	12.759	0.092	12.759	0.109	12.759	0.119	12.759	0.128	12.759	0.134

Table E.1

Maintenance activities for interstate and primary pavements under different categories based on VDOT Chowdhury (2016).

Activity Category	Activities
Do Nothing (DN)	N/A
Preventive Maintenance (PM)	1. Minor patching (<5% of pavement Area; surface patching; depth 2") 2. Crack sealing 3. Thin treatments (chip seal, slurry seal, latex, Thin Hotmix Asphalt Concrete (THMACO), 'Novachip' etc.)
Corrective Maintenance (CM)	1. Moderate patching (<10% of pavement area; partial depth patching; depth 6") 2. Partial depth patching (<10% of pavement area; depth 4"-6"); and surface treatment 3. Partial depth patching (<10% of pavement area; depth 4"-6"); and thin (≤ 2 ") AC overlay 4. ≤ 2 " Milling; and ≤ 2 " AC overlay
Restorative Maintenance (RM)	1. Heavy patching (20% of pavement area; full depth patching; depth 12") 2. ≤ 4 " Milling and replace with ≤ 4 " AC overlay 3. Full depth patching (<20% of pavement area; full depth patching; depth 9"-12"); and 4" AC overlay 4. Cold in-place recycling
Rehabilitation /Reconstruction (RC)	1. Mill, break and seat and 9"-12" AC overlay 2. Reconstruction 3. Full depth reclamation

Table E.2Observation probability $p(o_i|s_i)$ given actual CCI state s_i with $\sigma_{error}^2 = 72$.

Actual state	$p(o_i = 6 s_i)$	$p(o_i = 5 s_i)$	$p(o_i = 4 s_i)$	$p(o_i = 3 s_i)$	$p(o_i = 2 s_i)$	$p(o_i = 1 s_i)$
$s_i = 6$	0.687	0.259	0.054	0.000	0.000	0.000
$s_i = 5$	0.276	0.422	0.297	0.005	0.000	0.000
$s_i = 4$	0.023	0.139	0.648	0.167	0.022	0.001
$s_i = 3$	0.000	0.003	0.266	0.455	0.248	0.028
$s_i = 2$	0.000	0.000	0.031	0.224	0.486	0.259
$s_i = 1$	0.000	0.000	0.000	0.005	0.059	0.936

Table E.3Observation probability $p(o_i|s_i)$ given actual CCI state s_i with $\sigma_{error}^2 = 18$.

Actual state	$p(o_i = 6 s_i)$	$p(o_i = 5 s_i)$	$p(o_i = 4 s_i)$	$p(o_i = 3 s_i)$	$p(o_i = 2 s_i)$	$p(o_i = 1 s_i)$
$s_i = 6$	0.801	0.197	0.002	0.000	0.000	0.000
$s_i = 5$	0.153	0.664	0.183	0.000	0.000	0.000
$s_i = 4$	0.001	0.078	0.822	0.099	0.000	0.000
$s_i = 3$	0.000	0.000	0.149	0.693	0.158	0.000
$s_i = 2$	0.000	0.000	0.001	0.137	0.718	0.144
$s_i = 1$	0.000	0.000	0.000	0.000	0.042	0.958

Table E.4

Minor Repair transition probabilities for 6 CCI states.

Condition State	$s_{t+1} = 6$	$s_{t+1} = 5$	$s_{t+1} = 4$	$s_{t+1} = 3$	$s_{t+1} = 2$	$s_{t+1} = 1$
$s_t = 6$	0.97	0.03	0.00			
$s_t = 5$	0.87	0.10	0.03			
$s_t = 4$	0.40	0.47	0.10	0.03		
$s_t = 3$		0.40	0.47	0.10	0.03	
$s_t = 2$			0.40	0.47	0.10	0.03
$s_t = 1$				0.40	0.47	0.13
Deterioration rate	Does not change					

and *Reconstruction*) transition probabilities are shown in Tables E.4–E.6 for CCI states, and Tables E.7–E.9 for IRI states. Table E.12 shows the durations used for the completion of the maintenance actions per lane-mile of pavement. Similarly, Tables E.13–E.15 provide the action transition probabilities for the 7 deck states for maintenance actions *Minor Repair*, *Major Repair*, and *Reconstruction*, respectively; Tables E.16–E.17 provide the observation probabilities for the 7 deck states, for low- and high-fidelity inspection techniques, respectively, and, finally, Table E.18 reports the maintenance action durations in days for decks.

Table E.5

Major Repair transition probabilities for 6 CCI states.

Condition State	$s_{t+1} = 6$	$s_{t+1} = 5$	$s_{t+1} = 4$	$s_{t+1} = 3$	$s_{t+1} = 2$	$s_{t+1} = 1$
$s_t = 6$	1.00	0.00	0.00			
$s_t = 5$	0.96	0.04	0.00			
$s_t = 4$	0.80	0.20	0.00			
$s_t = 3$	0.65	0.25	0.10			
$s_t = 2$	0.50	0.30	0.20			
$s_t = 1$	0.40	0.30	0.30			
Deterioration rate	Reset by 5 years					

Table E.6

Reconstruction transition probabilities for 6 CCI states.

Condition State	$s_{t+1} = 6$	$s_{t+1} = 5$	$s_{t+1} = 4$	$s_{t+1} = 3$	$s_{t+1} = 2$	$s_{t+1} = 1$
$s_t = 6$	1.00	0.00	0.00			
$s_t = 5$	1.00	0.00	0.00			
$s_t = 4$	1.00	0.00	0.00			
$s_t = 3$	1.00	0.00	0.00			
$s_t = 2$	1.00	0.00	0.00			
$s_t = 1$	1.00	0.00	0.00			
Deterioration rate	Reset to newly built					

Table E.7

Minor Repair transition probabilities for 5 IRI states.

Condition State	$s_{t+1} = 5$	$s_{t+1} = 4$	$s_{t+1} = 3$	$s_{t+1} = 2$	$s_{t+1} = 1$
$s_t = 5$	0.97	0.03	0.00		
$s_t = 4$	0.85	0.12	0.03		
$s_t = 3$	0.45	0.40	0.12	0.03	
$s_t = 2$		0.45	0.40	0.12	0.03
$s_t = 1$			0.45	0.40	0.15

Table E.8

Major Repair transition probabilities for 5 IRI states.

Condition State	$s_{t+1} = 5$	$s_{t+1} = 4$	$s_{t+1} = 3$	$s_{t+1} = 2$	$s_{t+1} = 1$
$s_t = 5$	1.00	0.00	0.00		
$s_t = 4$	0.95	0.05	0.00		
$s_t = 3$	0.80	0.20	0.00		
$s_t = 2$	0.70	0.25	0.05		
$s_t = 1$	0.45	0.35	0.20		

Table E.9

Reconstruction transition probabilities for 5 IRI states.

Condition State	$s_{t+1} = 5$	$s_{t+1} = 4$	$s_{t+1} = 3$	$s_{t+1} = 2$	$s_{t+1} = 1$
$s_t = 5$	1.00	0.00	0.00		
$s_t = 4$	1.00	0.00	0.00		
$s_t = 3$	1.00	0.00	0.00		
$s_t = 2$	1.00	0.00	0.00		
$s_t = 1$	1.00	0.00	0.00		

Table E.10Inspection action costs (in USD/lane-mile or USD/m²) for CCI using 3 fidelity levels.

Inspection Technique	Description	Observation error (σ_{error}^2)	Cost (USD/lane-mile)	Cost (USD/m ²)
i_2	High fidelity	18	0.57	0.16
i_1	Low fidelity	72	0.27	0.08
i_0	No Inspection	∞	0.00	0.00

Table E.11Observation probability $p(o_i|s_t)$ for IRI given actual state s_t .

Inspection Technique	SD (m/km)	$p(o_i = j-2 s_t = j)$	$p(o_i = j-1 s_t = j)$	$p(o_i = j s_t = j)$	$p(o_i = j+1 s_t = j)$	$p(o_i = j+2 s_t = j)$	Cost (USD/m ²)
i_2	0.08	0	0.05	0.90	0.05	0	0.10
i_1	0.32	0	0.20	0.60	0.20	0	0.03
i_0	∞	0.20	0.20	0.20	0.20	0.20	0

Table E.12

Maintenance actions duration for pavements (in days per lane-mile).

Actions	Days per lane-mile	Additional days per mile for shoulder, etc.
Do Nothing	0	0
Minor Repair	3.5	1
Major Repair	6.5	2
Reconstruction	32	10

Table E.13

Minor Repair transition probabilities for deck states.

Condition State	$s_{t+1} = 9$	$s_{t+1} = 8$	$s_{t+1} = 7$	$s_{t+1} = 6$	$s_{t+1} = 5$	$s_{t+1} = 4, \dots$	$s_{t+1} = \text{failed}$
$s_t = 9$	0.97	0.03	0.00				
$s_t = 8$	0.85	0.12	0.03				
$s_t = 7$	0.40	0.45	0.12	0.03			
$s_t = 6$		0.40	0.45	0.12	0.03		
$s_t = 5$			0.40	0.45	0.12	0.03	
$s_t = 4, \dots$				0.40	0.45	0.15	
$s_t = \text{failed}$							1.0
Deterioration rate	Does not change						

Table E.14

Major Repair transition probabilities for deck states.

Condition State	$s_{t+1} = 9$	$s_{t+1} = 8$	$s_{t+1} = 7$	$s_{t+1} = 6$	$s_{t+1} = 5$	$s_{t+1} = 4, \dots$	$s_{t+1} = \text{failed}$
$s_t = 9$	1.00	0.00	0.00				
$s_t = 8$	0.95	0.05	0.00				
$s_t = 7$	0.80	0.20	0.00				
$s_t = 6$	0.60	0.30	0.10				
$s_t = 5$	0.40	0.40	0.20				
$s_t = 4, \dots$	0.30	0.40	0.30				
$s_t = \text{failed}$							1.0
Deterioration rate	Reset by 5 years						

Table E.15
Reconstruction transition probabilities for deck states.

Condition State	$s_{t+1} = 9$	$s_{t+1} = 8$	$s_{t+1} = 7$	$s_{t+1} = 6$	$s_{t+1} = 5$	$s_{t+1} = 4, \dots$	$s_{t+1} = \text{failed}$
$s_t = 9$	1.00						
$s_t = 8$	1.00						
$s_t = 7$	1.00						
$s_t = 6$	1.00						
$s_t = 5$	1.00						
$s_t = 4, \dots$	1.00						
$s_t = \text{failed}$	1.00						
Deterioration rate	Reset to newly built deck						

Table E.16
Observation probability $p(o_t | s_t)$ given state s_t , for low fidelity inspection techniques for decks.

Condition State	$p(o_t = 9 s_t)$	$p(o_t = 8 s_t)$	$p(o_t = 7 s_t)$	$p(o_t = 6 s_t)$	$p(o_t = 5 s_t)$	$p(o_t = 4, \dots s_t)$	$p(o_t = \text{failed} s_t)$
$s_t = 9$	0.80	0.15	0.05				
$s_t = 8$	0.15	0.65	0.15	0.05			
$s_t = 7$	0.05	0.15	0.60	0.15	0.05	0.0	
$s_t = 6$		0.05	0.15	0.60	0.15	0.05	
$s_t = 5$			0.05	0.15	0.65	0.15	
$s_t = 4, \dots$				0.05	0.15	0.80	
$s_t = \text{failed}$							1.0

Table E.17
Observation probability $p(o_t | s_t)$ given state s_t , for high fidelity inspection techniques for decks.

Condition State	$p(o_t = 9 s_t)$	$p(o_t = 8 s_t)$	$p(o_t = 7 s_t)$	$p(o_t = 6 s_t)$	$p(o_t = 5 s_t)$	$p(o_t = 4, \dots s_t)$	$p(o_t = \text{failed} s_t)$
$s_t = 9$	0.90	0.10					
$s_t = 8$	0.10	0.80	0.10				
$s_t = 7$		0.10	0.80	0.10			
$s_t = 6$		0.05	0.10	0.80	0.10		
$s_t = 5$				0.10	0.80	0.10	
$s_t = 4, \dots$					0.10	0.90	
$s_t = \text{failed}$							1.0

Table E.18
Deck maintenance action durations in days.

Actions	Type I Bridges	Type II Bridges	Type III Bridges
Do Nothing	0	0	0
Minor Repair	25	12	6
Major Repair	70	30	15
Reconstruction	300	150	70

Data availability

All data used in this work are reported in the manuscript.

References

- ADOT. (2018). *Production rates guidelines for Arizona highway construction*. <https://azdot.gov/sites/default/files/2019/06/adot-typical-production-rates.pdf> (Last accessed 23 February 2026).
- American Road and Transportation Builders Association. (2026). <https://www.artba.org/> (Last accessed 23 February 2026).
- American Society of Civil Engineers. (2025). *2025 infrastructure report card*. tech. rep, American Society of Civil Engineers.: <https://infrastructurereportcard.org/wp-content/uploads/2025/03/Full-Report-2025-Natl-IRC-WEB.pdf> (Last accessed 23 February 2026).
- Andriotis, C. P., & Papakonstantinou, K. G. (2019). Managing engineering systems with large state and action spaces through deep reinforcement learning. *Reliability Engineering & System Safety*, 191, 106483.
- Andriotis, C. P., & Papakonstantinou, K. G. (2019). Life-cycle policies for large engineering systems under complete and partial observability. In *13th international conference on applications of statistics and probability in civil engineering (ICASP)*, Seoul, South Korea.
- Andriotis, C. P., & Papakonstantinou, K. G. (2021). Deep reinforcement learning driven inspection and maintenance planning under incomplete information and constraints. *Reliability Engineering & System Safety*, 212, 107551.
- Andriotis, C. P., & Papakonstantinou, K. G. (2022). Optimizing policies for deteriorating systems considering ordered action structuring and value of information. In *13th international conference on structural safety & reliability (ICOSAR)*.
- Arcieri, G., Hoelzl, C., Schwery, O., Straub, D., Papakonstantinou, K. G., & Chatzi, E. (2023). Bridging POMDPs and Bayesian decision making for robust maintenance planning under model uncertainty: An application to railway systems. *Reliability Engineering & System Safety*, 239, 109496.
- Arcieri, G., Hoelzl, C., Schwery, O., Straub, D., Papakonstantinou, K. G., & Chatzi, E. (2024). POMDP inference and robust solution via deep reinforcement learning: An application to railway optimal maintenance. *Machine Learning*, 113(10), 7967–7995.
- Arcieri, G., Duthé, G., Muller, C., Haener, D., Papakonstantinou, K. G., Straub, D., & Chatzi, E. (2025). Graph-based multi-agent reinforcement learning for railway infrastructure decision support. In *14th international conference on structural safety and Reliability-ICOSAR'25*. Scipedia.
- Azad, N., Hassini, E., & Verma, M. (2016). Disruption risk management in railroad networks: An optimization-based methodology and a case study. *Transportation Research Part B: Methodological*, 85, 70–88.
- Bellman, R. (1957). *Dynamic programming*. Princeton University Press.
- Bernstein, D. S., Givan, R., Immerman, N., & Zilberstein, S. (2002). The complexity of decentralized control of Markov decision processes. *Mathematics of Operations Research*, 27(4), 819–840.
- Bertsekas, D. P. (1997). Nonlinear programming. *Journal of the Operational Research Society*, 48(3), 334.
- Bhattacharya, A., Saifullah, M., & Papakonstantinou, K. G. (2025). District-level bridge networks management with multi-agent reinforcement learning: From theory to real-world application. *Structure and Infrastructure Engineering*, 21(11–12), 2064–2082.
- Bhustali, P., Morato, P. G., Papakonstantinou, K. G., & Andriotis, C. P. (2026). *The price of decentralization in managing engineering systems through multi-agent reinforcement learning*. arXiv preprint arXiv:2603.11884.
- Bismut, E., & Straub, D. (2021). Optimal adaptive inspection and maintenance planning for deteriorating structural systems. *Reliability Engineering & System Safety*, 215, 107891.
- Bocchini, P., & Frangopol, D. (2011). A probabilistic computational framework for bridge network optimal maintenance scheduling. *Reliability Engineering & System Safety*, 96(2), 332–349.
- Chang, Y., Liu, R., & Tang, Y. (2023). Segment-condition-based railway track maintenance schedule optimization. *Computer-Aided Civil and Infrastructure Engineering*, 38(2), 160–193.
- Chowdhury, T. (2016). *Supporting document for the development and enhancement of the pavement maintenance decision matrices used in the needs-based analysis*. tech. rep, Virginia Transportation Research Council, Maintenance Division, Richmond, VA.
- Chu, J. C., & Chen, Y.-J. (2012). Optimal threshold-based network-level transportation infrastructure life-cycle management with heterogeneous maintenance actions. *Transportation Research Part B: Methodological*, 46(9), 1123–1143.
- Cost per mile models by FDOT. (2022). <https://www.fdot.gov/> (Last accessed 23 February 2026).
- Daily vehicle miles traveled (DVMT). (2020). http://166.67.200.218/info/resources/Traffic_2020/VMTReport_2100_2020.pdf (Last accessed 23 February 2026).
- Degrís, T., White, M., & Sutton, R. S. (2012). arXiv preprint *Off-policy actor-critic*. arXiv:1205.4839.
- Do, P., Nguyen, V.-T., Voisin, A., Jung, B., & Neto, W. A. F. (2024). Multi-agent deep reinforcement learning-based maintenance optimization for multi-dependent component systems. *Expert Systems with Applications*, 245, 123144.
- Du, A., & Ghavidel, A. (2022). Parameterized deep reinforcement learning-enabled maintenance decision-support and life-cycle risk assessment for highway bridge portfolios. *Structural Safety*, 97, 102221.
- Dye Management Group Inc. (2014). *Monitoring highway assets with remote technology*. Technical Report, Michigan Department of Transportation.
- Faddoul, R., Raphael, W., Soubra, A.-H., & Chateaufort, A. (2013). Incorporating Bayesian networks in Markov decision processes. *Journal of Infrastructure Systems*, 19(4), 415–424.
- Fard, A., & Yuan, A. X. X. (2025). arXiv preprint *Multi-year maintenance planning for large-scale infrastructure systems: A novel network deep Q-learning approach* arXiv:2507.18732.
- Federal Highway Administration and United States Department of Transportation. (2026a). 23 CFR 490.411: Establishment of minimum level for condition for bridges. <https://www.ecfr.gov/current/title-23/section-490.411> (Last accessed 23 February 2026).
- Federal Highway Administration and United States Department of Transportation. (2026b). 23 CFR 490.315: Establishment of minimum level for condition of pavements. <https://www.ecfr.gov/current/title-23/part-490/section-490.315> (Last accessed 23 February 2026).
- Federal Highway Administration. (2019). *Bridge replacement unit costs*. Washington D.C: FHWA.
- Federal Highway Administration. (2020). Work zone road user costs - concepts and applications: Chapter 2. work zone road user costs. <https://ops.fhwa.dot.gov/wz/resources/publications/fhwahop12005/Section2.htm> (Last accessed 23 February 2026).
- Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., & Whiteson, S. (2018). Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32).
- Frangopol, D. M., & Liu, M. (2007). Maintenance and management of civil infrastructure based on condition, safety, optimization, and life-cycle cost. *Structure and Infrastructure Engineering*, 3(1), 29–41.
- Frangopol, D. M., & Liu, M. (2007). Bridge network maintenance optimization using stochastic dynamic programming. *Journal of Structural Engineering*, 133(12), 1772–1782.
- Gillespie, T. D., Paterson, W., & Sayers, M. W. (1986). *Guidelines for conducting and calibrating road roughness measurements*. tech. rep, The World Bank.
- Goodspeed, C., & Brown, J. (2017). *Gilford rapid bridge deck replacement, a project summary*. tech. rep, University of New Hampshire.: <https://mm.nh.gov/files/uploads/dot/remote-docs/nhdot-research-15680x-2017-summary-report.pdf> (Last accessed 23 February 2026).
- Gronauer, S., & Diepold, K. (2022). Multi-agent deep reinforcement learning: A survey. *Artificial Intelligence Review*, 1–49.
- Guillaumot, V. M., Durango-Cohen, P. L., & Madanat, S. M. (2003). Adaptive optimization of infrastructure maintenance and inspection decisions under performance model uncertainty. *Journal of Infrastructure Systems*, 9(4), 133–139.
- Hamida, Z., & Goulet, J.-A. (2023). Hierarchical reinforcement learning for transportation infrastructure maintenance planning. *Reliability Engineering & System Safety*, 235, 109214.
- Hampton Roads Planning District Commission (HRPDC). (2026). Hampton roads maps. <https://www.hrpdcva.gov/1009/Data-Maps> (Last accessed 23 February 2026).
- Hampton Roads Transportation Planning Organization (HRTPO). (2024). *The state of transportation in Hampton roads*. <https://hrpdcva.gov/DocumentCenter/View/13727/T25-01-State-of-Transportation-2024-Final-PDF?bidId=> (Last accessed 23 February 2026).
- Han, X., & Frangopol, D. M. (2022). Risk-informed bridge optimal maintenance strategy considering target service life and user cost at project and network levels. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 8(4), 04022050.
- Han, X., & Frangopol, D. M. (2023). Life-cycle risk-based optimal maintenance strategy for bridge networks subjected to corrosion and seismic hazards. *Journal of Bridge Engineering*, 28(1), 04022128.
- Hasnat, M., Singh, R. R., Kutay, M. E., Bryce, J., Haider, S. W., & Cetin, B. (2023). Comparative study of different condition indices using Michigan department of transportation's flexible distress data. *Transportation Research Record*, 03611981231156917.
- Kaelbling, L., Littman, M., & Cassandra, A. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1), 99–134.
- Katcha, S. W., Ercisli, S., Flintsch, G. W., Bryce, J. M., & Diefenderfer, B. K. (2016). *Development of enhanced pavement deterioration curves*. tech. rep, Virginia Transportation Research Council.
- Konda, V. R., & Tsitsiklis, J. N. (2000). Actor-critic algorithms. In *Advances in neural information processing systems* (pp. 1008–1014).
- Kurniawati, H., Hsu, D., & Lee, W. S. (2008). Sarsop: Efficient point-based POMDP planning by approximating optimally reachable belief spaces. In *Robotics: Science and systems* (Vol. 2008). Citeseer.
- Lai, L., Dong, Y., Andriotis, C. P., Wang, A., & Lei, X. (2024). Synergetic-informed deep reinforcement learning for sustainable management of transportation networks with large action spaces. *Automation in Construction*, 160, 105302.
- Lee, J., & Mitici, M. (2023). Deep reinforcement learning for predictive aircraft maintenance using probabilistic remaining-useful-life prognostics. *Reliability Engineering & System Safety*, 230, 108908.
- Lee, D., & Song, J. (2023). Risk-informed operation and maintenance of complex life-line systems using parallelized multi-agent deep Q-Network. *Reliability Engineering & System Safety*, 109512.
- Lei, X., Xia, Y., Deng, L., & Sun, L. (2022). A deep reinforcement learning framework for life-cycle maintenance planning of regional deteriorating bridges using inspection data. *Structural and Multidisciplinary Optimization*, 65(5), 149.
- Lei, X., Dong, Y., & Frangopol, D. M. (2023). Sustainable life-cycle maintenance policy-making for network-level deteriorating bridges with a convolutional autoencoder-structured reinforcement learning agent. *Journal of Bridge Engineering*, 28(9), 04023063.
- Leroy, P., Morato, P. G., Pisane, J., Kolios, A., & Ernst, D. (2023). *IMP-MARL: A suite of environments for large-scale infrastructure management planning via MARL*. arXiv preprint arXiv:2306.11551.
- Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2016). Continuous control with deep reinforcement learning. In *Proceedings of the international conference on learning representations (ICLR)*.

- Lovejoy, W. S. (1991). A survey of algorithmic methods for partially observed Markov decision processes. *Annals of Operations Research*, 28(1), 47–65.
- Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., & Mordatch, I. (2017). *Multi-agent actor-critic for mixed cooperative-competitive environments*. arXiv preprint arXiv:1706.02275.
- Madanat, S., & Ben-Akiva, M. (1994). Optimal inspection and repair policies for infrastructure. *Transportation Science*, 28(1), 55–62.
- Madanat, S. (1993). Optimal infrastructure management decisions under uncertainty. *Transportation Research Part C: Emerging Technologies*, 1(1), 77–88.
- Manafpour, A., Guler, I., Radlińska, A., Rajabipour, F., & Warn, G. (2018). Stochastic analysis and time-based modeling of concrete bridge deck deterioration. *Journal of Bridge Engineering*, 23(9), 04018066.
- Memarzadeh, M., & Pozzi, M. (2016). Integrated inspection scheduling and maintenance planning for infrastructure systems. *Computer-Aided Civil and Infrastructure Engineering*, 31(6), 403–415.
- Memarzadeh, M., Pozzi, M., & Zico Kolter, J. (2015). Optimal planning and learning in uncertain environments for the management of wind farms. *Journal of Computing in Civil Engineering*, 29(5), 04014076.
- Mendoza, J., Bismut, E., Straub, D., & Köhler, J. (2021). Risk-based fatigue design considering inspections and maintenance. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 7(1), 04020055.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fiedelnd, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529–533.
- Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., & Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning* (pp. 1928–1937). PMLR.
- Mohammadi, R., & He, Q. (2022). A deep reinforcement learning approach for rail renewal and maintenance planning. *Reliability Engineering & System Safety*, 225, 108615.
- Morato, P. G., Papakonstantinou, K. G., Andriotis, C. P., Nielsen, J. S., & Rigo, P. (2022). Optimal inspection and maintenance planning for deteriorating structural components through dynamic Bayesian networks and Markov decision processes. *Structural Safety*, 94, 102140.
- Morato, P. G., Andriotis, C. P., Papakonstantinou, K. G., & Rigo, P. (2023). Inference and dynamic decision-making for deteriorating systems with probabilistic dependencies through Bayesian networks and deep reinforcement learning. *Reliability Engineering & System Safety*, 235, 109144.
- Nguyen, K.-A., Do, P., & Grall, A. (2015). Multi-level predictive maintenance for multi-component systems. *Reliability Engineering & System Safety*, 144, 83–94.
- Nguyen, V.-T., Do, P., Voisin, A., & Iung, B. (2022). Weighted-QMIX-based optimization for maintenance decision-making of multi-component systems. In *PHM Society European Conference* (Vol. 7, pp. 360–367).
- Nguyen, V.-T., Do, P., Vosin, A., & Iung, B. (2022). Artificial-intelligence-based maintenance decision-making and optimization for multi-state component systems. *Reliability Engineering & System Safety*, 228, 108757.
- Oakgrove. (2013). *Accelerated bridge construction-four bridge deck replacements in region 5*. Oakgrove Construction, Inc.
- Oliehoek, F. A., & Amato, C. (2016). *A concise introduction to decentralized POMDPs*. Springer.
- Oliehoek, F. A., Spaan, M. T., & Vlassis, N. (2008). Optimal and approximate Q-value functions for decentralized POMDPs. *Journal of Artificial Intelligence Research*, 32, 289–353.
- Pan, Y., Qin, J., Zhang, L., Pan, W., & Chen, J.-J. (2024). A probabilistic deep reinforcement learning approach for optimal monitoring of a building adjacent to deep excavation. *Computer-Aided Civil and Infrastructure Engineering*, 39(5), 656–678.
- Papakonstantinou, K. G., Eshra, E., & Nikbakht, H. (2023). Hamiltonian MCMC based framework for time-variant rare event uncertainty quantification. In *14th international conference on applications of statistics and probability in civil engineering (ICASP)*, Dublin, Ireland.
- Papakonstantinou, K. G., Nikbakht, H., & Eshra, E. (2023). Hamiltonian MCMC methods for estimating rare event probabilities in high-dimensional problems. *Probabilistic Engineering Mechanics*, 74, 103485.
- Papakonstantinou, K. G., & Shinozuka, M. (2014). Optimum inspection and maintenance policies for corroded structures using partially observable Markov decision processes and stochastic, physically based models. *Probabilistic Engineering Mechanics*, 37, 93–108.
- Papakonstantinou, K. G., & Shinozuka, M. (2014). Planning structural inspection and maintenance policies via dynamic programming and Markov processes. Part I: Theory. *Reliability Engineering & System Safety*, 130, 202–213.
- Papakonstantinou, K. G., & Shinozuka, M. (2014). Planning structural inspection and maintenance policies via dynamic programming and Markov processes. Part II: Pomdp implementation. *Reliability Engineering & System Safety*, 130, 214–224.
- Papakonstantinou, K. G., Andriotis, C. P., & Shinozuka, M. (2016). Point-based POMDP solvers for life-cycle cost minimization of deteriorating structures. In *Proceedings of the 5th international symposium on life-cycle civil engineering (IALCCE 2016)*, Delft, The Netherlands (pp. 427).
- Papakonstantinou, K. G., Andriotis, C. P., & Shinozuka, M. (2018). POMDP and MOMDP solutions for structural life-cycle cost minimization under partial and mixed observability. *Structure and Infrastructure Engineering*, 14(7), 869–882.
- PennDOT. (2017). *Road maintenance and preservation (MaP)*. <https://www.pa.gov/content/dam/copapwp-pagov/en/pennidot/documents/about-us/documents/pennidot%20road%20initiative.pdf> (Last accessed 23 February 2026).
- PennDOT. (2019). *Publication 242 pavement policy manual*. (May 2015 Edition) Change No. 5, tech. rep, Pennsylvania Department of Transportation.: <https://www.butlercountypa.gov/DocumentCenter/View/1555/PennDOT-Publication-242-Pavement-Policy-Manual-PDF> (Last accessed 23 February 2026).
- Pennsylvania Department of Transportation and Bureau of Design. (2022). *Bridge management system 2 (BMS2) coding manual: Publication 100A*, 2022 edition.
- Petchrompo, S., & Parlikad, A. K. (2019). A review of asset management literature on multi-asset systems. *Reliability Engineering & System Safety*, 181, 181–201.
- Pineau, J., Gordon, G., & Thrun, S. (2003). Point-based value iteration: An anytime algorithm for POMDPs. In *IJCAI* (Vol. 3, pp. 1025–1032).
- Rashid, T., Samvelyan, M., Schroeder, C., Farquhar, G., Foerster, J., & Whiteson, S. (2018). QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *International conference on machine learning* (pp. 4295–4304). PMLR.
- Rashid, T., Farquhar, G., Peng, B., & Whiteson, S. (2020). Weighted QMIX: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 33, 10199–10210.
- Rodríguez, M. L. R., Kubler, S., de Giorgio, A., Cordy, M., Robert, J., & Le Traon, Y. (2022). Multi-agent deep reinforcement learning based predictive maintenance on parallel machines. *Robotics and Computer-Integrated Manufacturing*, 78, 102406.
- Sánchez-Silva, M., Frangopol, D. M., Padgett, J., & Soliman, M. (2016). Maintenance and operation of infrastructure systems. *Journal of Structural Engineering*, 142(9), F4016004.
- Saifullah, M., Andriotis, C. P., & Papakonstantinou, K. G. (2023). The role of value of information in multi-agent deep reinforcement learning for optimal decision-making under uncertainty. In *14th international conference on applications of statistics and probability in civil engineering (ICASP14)*, Dublin.
- Saydam, D., & Frangopol, D. M. (2015). Risk-based maintenance optimization of deteriorating bridges. *Journal of Structural Engineering*, 141(4), 04014120.
- Schöbi, R., & Chatzi, E. N. (2016). Maintenance planning using continuous-state partially observable Markov decision processes and non-linear action models. *Structure and Infrastructure Engineering*, 12(8), 977–994.
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). *Proximal policy optimization algorithms*. arXiv preprint arXiv:1707.06347.
- Shani, G., Pineau, J., & Kaplow, R. (2013). A survey of point-based POMDP solvers. *Autonomous Agents and Multi-Agent Systems*, 27(1), 1–51.
- Shon, H., & Lee, J. (2021). Integrating multi-scale inspection, maintenance, rehabilitation, and reconstruction decisions into system-level pavement management systems. *Transportation Research Part C: Emerging Technologies*, 131, 103328.
- Sondik, E. (1971). *The optimal control of partially observable Markov processes*. Stanford, CA: Stanford University, Stanford Electronics Labs.
- Spaan, M., & Vlassis, N. (2005). Perseus: Randomized point-based value iteration for POMDPs. *Journal of Artificial Intelligence Research*, 24, 195–220.
- Sresakoolchai, J., & Kaewwunruen, S. (2023). Railway infrastructure maintenance efficiency improvement using deep reinforcement learning integrated with digital twin based on track geometry and component defects. *Scientific Reports*, 13(1), 2439.
- Su, J., Adams, S., & Beling, P. (2021). Value-decomposition multi-agent actor-critics. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 35, pp. 11352–11360).
- Su, J., Huang, J., Adams, S., Chang, Q., & Beling, P. A. (2022). Deep multi-agent reinforcement learning for multi-level preventive maintenance in manufacturing systems. *Expert Systems with Applications*, 192, 116323.
- Sunehag, P., Lever, G., Gruslys, A., Czarniecki, W. M., Zambaldi, V., Jaderberg, M., Lanctot, M., Sonnerat, N., Leibo, J. Z., Tuyls, K., & Graepel, T. (2017). *Value-decomposition networks for cooperative multi-agent learning*. arXiv preprint arXiv:1706.05296.
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.
- Sutton, R. S., McAllester, D. A., Singh, S. P., & Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems* (pp. 1057–1063).
- Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, 3(1), 9–44.
- United States Department of Transportation. (2000). 1999 status of the nation's highways, bridges and transit: Conditions & performance, a report to congress: Executive summary. <https://rosap.nhtl.bts.gov/view/dot/15379> (Last accessed 23 February 2026).
- Vadapat, G., Stoffels, S., & Dixon, K. (2000). Road user cost models for network-level pavement management. *Transportation Research Record*, 1699(1), 49–57.
- Van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double Q-learning. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 30).
- VDOT. (2022). *Preservation, maintenance, repair, widening and rehabilitation, Chapter 32*. tech. rep, Virginia Department of Transportation.: <https://www.virginia.gov/business/resources/bridge/manuals/part2/chapter32.pdf> (Last accessed 23 February 2026).
- Virginia Department of Transportation. (2018). *Approval of the report on the Commonwealth's large and unique bridge and tunnel structures: Vital infrastructure*. <https://ctb.virginia.gov/media/ctb/agendas-and-meeting-minutes/2018/dec/res/14title-approval-of-the-report-on-the-commonwealths-large-and-unique-bridge-and-tunnel-structures-vital-infrastructure.pdf> (Last accessed 23 February 2026).
- Virginia Department of Transportation. (2019). *Maintenance and operations comprehensive review*. https://www.vdot.virginia.gov/media/vdotvirginiagov/about/legislative-studies-and-reports/Maintenance_and_Operations_Comprehensive_Review_2019.pdf (Last accessed 23 February 2026).
- Virginia Department of Transportation. Pavement condition map.: <https://vdot.maps.arcgis.com/apps/webappviewer/index.html?id=d93cecc94efb4bda9213d0c3cef73ce4> (Last accessed 23 February 2026).
- Virginia Department of Transportation. Structure condition map.: <https://dashboard.virginiadot.org/pages/maintenance/bridge.aspx> (Last accessed 23 February 2026).
- Wang, X., Wang, H., & Qi, C. (2016). Multi-agent reinforcement learning based maintenance policy for a resource constrained flow line system. *Journal of Intelligent Manufacturing*, 27, 325–333.
- Wang, Z., Bapst, V., Heess, N., Munos, R., Kavukcuoglu, K., & Freitas, N. (2016). arXiv preprint *Sample efficient actor-critic with experience replay* arXiv:1611.01224.
- Wells, D. T. (1995). *Technical assistance report: Maintenance, repair, and rehabilitation unit costs for PONTIS*. Tech. rep, Virginia Transportation Research Council. <https://rosap.nhtl.bts.gov/view/dot/20088> (Last accessed 23 February 2026).

- Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4), 229–256.
- Yang, D. Y., & Frangopol, D. M. (2019). Life-cycle management of deteriorating civil infrastructure considering resilience to lifetime hazards: A general approach based on renewal-reward processes. *Reliability Engineering & System Safety*, 183, 197–212.
- Yang, D. Y., & Frangopol, D. M. (2021). Risk-based inspection planning of deteriorating structures. *Structure and Infrastructure Engineering*, 18(1), 109–128.
- Yang, D. Y. (2022). Deep reinforcement learning-enabled bridge management considering asset and network risks. *Journal of Infrastructure Systems*, 28(3), 04022023.
- Yang, D. Y. (2022). Adaptive risk-based life-cycle management for large-scale structures using deep reinforcement learning and surrogate modeling. *Journal of Engineering Mechanics*, 148(1), 04021126.
- Yao, L., Dong, Q., Jiang, J., & Ni, F. (2020). Deep reinforcement learning for long-term pavement maintenance planning. *Computer-Aided Civil and Infrastructure Engineering*, 35(11), 1230–1245.
- Yao, L., Leng, Z., Jiang, J., & Ni, F. (2024). A multi-agent reinforcement learning model for maintenance optimization of interdependent highway pavement networks. *Computer-Aided Civil and Infrastructure Engineering*, 39(19), 2951–2970.
- Zhang, E., Zhang, R., & Masoud, N. (2023). Predictive trajectory planning for autonomous vehicles at intersections using reinforcement learning. *Transportation Research Part C: Emerging Technologies*, 149, 104063.
- Zhang, J., Li, X., Yuan, Y., Yang, D., Xu, P., & Au, F. T. (2024). A multi-agent ranking proximal policy optimization framework for bridge network life-cycle maintenance decision-making. *Structural and Multidisciplinary Optimization*, 67(11), 194.
- Zhang, L., Gu, W., Byon, Y.-J., & Lee, J. (2023). Condition-based pavement management systems accounting for model uncertainty and facility heterogeneity with belief updates. *Transportation Research Part C: Emerging Technologies*, 148, 104054.
- Zhou, W., Miller-Hooks, E., Papakonstantinou, K. G., Stoffels, S., & McNeil, S. (2022). A reinforcement learning method for multi-asset roadway improvement scheduling considering traffic impacts. *Journal of Infrastructure Systems*, 28(4), 04022033.