Wind Energ. Sci., 11, 1185–1204, 2026
https://doi.org/10.5194/wes-11-1185-2026

eawe european academy of wind energy
Open Access WIND ENERGY SCIENCE

Review of deep reinforcement learning for offshore wind farm maintenance planning

Marco Borsotti, Xiaoli Jiang, and Rudy R. Negenborn

Department of Maritime & Transport Technology, Delft University of Technology, Delft, the Netherlands

Correspondence: Marco Borsotti (m.borsotti@tudelft.nl)

Received: 28 October 2025 – Discussion started: 7 November 2025
Revised: 13 March 2026 – Accepted: 16 March 2026 – Published: 13 April 2026

Abstract. Offshore wind farms face unique challenges in maintenance due to harsh weather, remote locations, and complex logistics. Traditional maintenance strategies often fail to optimize operations, leading to unplanned failures or unnecessary servicing. In recent years, deep reinforcement learning (DRL) has shown clear potential to tackle these challenges through a data-driven approach. This paper provides a critical review of representative DRL models for offshore wind farm maintenance planning, elaborating on both single- and multi-agent frameworks, diverse training algorithms, various problem formulations, and the integration of domain-specific knowledge. The review compares the benefits and limitations of these methods, identifying a significant gap in the widely adopted use of simplistic binary maintenance decisions, rather than including multi-level or imperfect repairs in the action space. In conclusion, this work suggests directions for future research to overcome current limitations and enhance the applicability of DRL methods in offshore wind maintenance.

1 Introduction

Offshore wind farm maintenance presents unique challenges due to harsh weather conditions, remote locations, and the intricate coordination of logistical resources (National Renewable Energy Laboratory, 2022). Storms, high winds, and unpredictable sea states create uncertainty in scheduling, while the limited accessibility of offshore sites further complicates intervention planning. Additionally, the need to allocate maintenance crews, vessels, and spare parts increases operational complexity, often leading to delays and higher costs. Maintenance strategies, such as corrective or scheduled preventive measures, struggle to mitigate these challenges, often resulting in unplanned failures or unnecessary servicing (Fox et al., 2022).

Traditionally, offshore wind O&M has been supported by deterministic or stochastic optimization models, rule-based policies, and predictive maintenance strategies informed by condition-monitoring and SCADA data. However, these approaches typically require predefined decision rules or restrictive modelling assumptions, limiting their ability to adapt to the dynamic and uncertain offshore environment (Borsotti et al., 2026).

In recent years, DRL has shown promising results as a data-driven approach to tackle these challenges. DRL is a class of algorithms that combine the sequential decision-making framework of reinforcement learning (RL) with the representational power of deep neural networks. In a typical RL setting, an agent interacts with an environment defined as a Markov decision process (MDP), observing a state s_t, selecting an action a_t according to a policy $\pi(a_t|s_t)$, and receiving a scalar reward r_t. The objective is to learn a policy that maximizes the expected cumulative discounted return $E_\pi[\sum_t \gamma^t r_t]$, where $\gamma \in [0, 1]$ is the discount factor determining how future rewards are weighted relative to immediate ones (Sutton and Barto, 2018). Classical RL struggles when the state or action space is high-dimensional or continuous, as in offshore wind maintenance, where turbine states, weather, and logistics create vast combinations. DRL overcomes this limitation by using deep networks to approximate policy and value functions, enabling end-to-end learning directly from high-dimensional or partially observed inputs such as condition monitoring or weather data. These features make DRL particularly suitable for complex, stochastic decision problems in offshore wind O&M, where the agent

must learn adaptive, long-horizon maintenance policies under uncertainty.

Figure 1 summarizes four recurring challenges in offshore wind O&M – harsh and uncertain weather, unplanned failures, remote locations with limited accessibility, and complex logistics requiring resource coordination – and highlights four corresponding opportunities where data-driven decision support, and DRL in particular, can add value. First, *adaptive decision-making and learning* directly target uncertainty: by learning policies that update actions based on newly observed information (e.g. condition-monitoring signals or revised weather forecasts), DRL can move beyond static decision rules and react to changing offshore conditions. Second, *proactive scheduling and resource allocation* address all challenges by exploiting forecasts and operational data to time interventions when access windows are likely (e.g. using metocean forecasts and, where available, operational sensing such as lidar-informed wind estimates) and to prioritize tasks before expected inaccessibility or risk escalation, thereby reducing weather-driven waiting time and avoidable downtime. Third, *data-driven optimization* provides a mechanism to jointly trade off competing objectives (e.g. cost, energy yield, risk, availability) under uncertainty, which is essential when failures and maintenance actions have long-horizon consequences. Finally, *scalable frameworks* respond to the growth in decision complexity as wind farms scale (more turbines, more components, more interacting constraints): function approximation, state abstraction (e.g. spatial or graph representations), and decentralized or hierarchical extensions provide pathways to maintain tractability when the state and decision spaces expand, increasing the complexity of logistics and resource coordination.

Despite these advantages, DRL also comes with well-recognized limitations that are particularly relevant for safety-critical infrastructure such as offshore wind. First, policies are usually represented by deep neural networks, which behave as black-box models and make it difficult for operators to understand or audit the rationale behind individual decisions; this lack of transparency is identified as a barrier to deployment and has motivated a dedicated line of work on explainable and safe reinforcement learning (Qing et al., 2022; Bui and Hollweg, 2024). Second, state-of-the-art DRL algorithms tend to be data- and computation-hungry, often requiring millions of interactions with an environment for training; this sample inefficiency and dependence on high-fidelity simulators are highlighted as key obstacles to applying DRL in real-world systems where experiments are costly or risky (Dulac-Arnold et al., 2021). Third, recent reviews of RL in power and energy systems emphasize the challenges of transferring policies from simulations to real assets, enforcing strict safety and reliability constraints, and encoding operational limits and human oversight in reward functions, all of which have so far limited large-scale industrial adoption (Pesántez et al., 2024; Bui and Hollweg, 2024).

A substantial body of literature has reviewed different aspects of data-driven O&M, but none provides a dedicated synthesis of deep reinforcement learning (DRL) for maintenance planning. For instance, Fox et al. (2022) review predictive and prescriptive O&M strategies, highlighting how data-driven prognostics can support maintenance planning but without examining learning-based sequential decision frameworks. Tusar and Sarker (2022) provide a systematic review of maintenance cost minimization models for offshore wind farms, focusing on optimization formulations and cost structures rather than on adaptive control or online decision-making. Similarly, reviews on machine-learning-based condition monitoring and prognostics, such as Stetco et al. (2019), Tautz-Weinert and Watson (2017), Pandit and Wang (2024), and Wang et al. (2026), synthesize diagnostic and remaining useful life (RUL) estimation techniques but do not address how such prognostic information interacts with maintenance-scheduling policies.

Reinforcement learning itself has also been surveyed within the wind and power-system domain. Narayanan (2023) reviews reinforcement-learning applications in wind energy, primarily in the context of control, forecasting, and wake steering. Abkar et al. (2023) survey RL approaches for wind farm flow control, while Li et al. (2023) discuss DRL for modern power-system control problems, including frequency and voltage regulation. Yet, none of these reviews analyse O&M decision processes, nor do they compare DRL architectures, modelling assumptions, or their integration with offshore wind maintenance requirements.

To the best of our knowledge, this is the first review that focuses specifically on DRL for offshore wind farm maintenance planning. In contrast to prior reviews, this work provides a focused synthesis of DRL approaches that have been proposed for, or are transferable to, offshore wind farm maintenance planning. Specifically, we (i) compare single- and multi-agent DRL frameworks and the algorithms they employ (value based, policy gradient, and actor critic) in relation to the maintenance decision problems they target; (ii) analyse the problem formulations adopted in these studies, including MDP, POMDP, graph-based, and hierarchical representations, and how they encode uncertainty, PHM information, wake effects, weather, and logistical constraints; and (iii) discuss the role of domain-specific knowledge and the remaining modelling gaps, with particular emphasis on the prevailing use of binary repair decisions instead of more realistic multi-level maintenance actions. By organizing the review along these dimensions, we aim to clarify what current DRL models can and cannot do for offshore wind maintenance planning and to outline research directions needed to move from promising simulation results toward practical adoption in real offshore wind operations.

Rather than attempting an exhaustive survey of all works in the DRL domain, we have deliberately narrowed our focus to a select group of key studies that exemplify the state-of-the-art in this area.

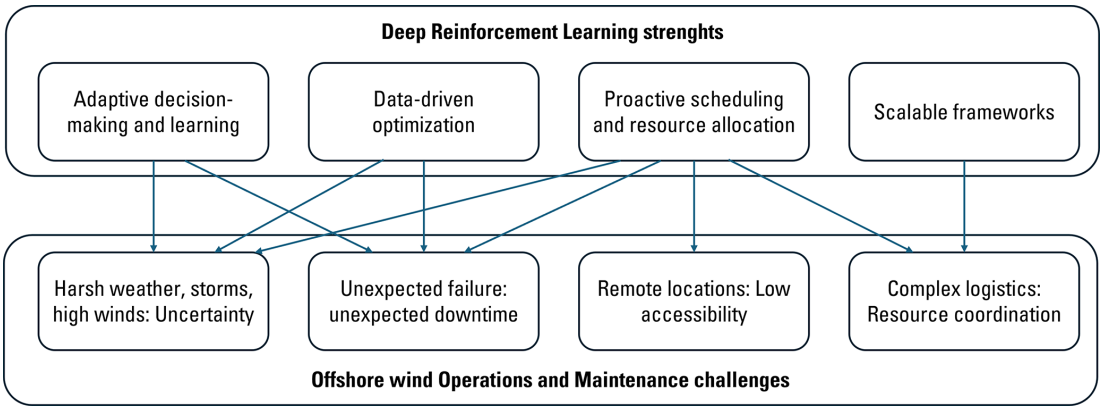

Figure 1. Challenges of offshore wind maintenance and opportunities for deep reinforcement learning methods.

To ensure transparency and reproducibility, the corpus analysed in this review was identified through a structured search process. Searches were conducted in Scopus, the Web of Science, ScienceDirect, Researchgate, and Google Scholar using combinations of the following terms: "offshore wind", "maintenance", "reinforcement learning", "deep reinforcement learning", "predictive maintenance", and "O&M optimization". The search window covered publications up to 2025.

Inclusion criteria were (i) studies proposing or evaluating DRL-based methods for maintenance or inspection planning of offshore or onshore wind systems, (ii) papers applying DRL to components or systems representative of wind turbine O&M, and (iii) articles providing sufficient methodological detail to characterize the learning formulation. Exclusion criteria included purely theoretical RL work, asset-management models without learning components, and high-level conceptual papers.

This protocol yielded a total of 54 papers after full-text screening. The term "deliberately narrowed" refers to the intentional focus on works with explicit DRL formulations for maintenance decision-making, excluding broader O&M optimization domains (e.g. dispatch, routing, or power forecasting) unless they directly informed maintenance planning.

The primary goal of this paper is to synthesize the advancements demonstrated by these representative models, critically assessing their strengths and identifying the limitations that still exist.

By doing so, we aim to demonstrate how DRL can effectively address the challenges of offshore wind maintenance planning, while also pinpointing areas that require further refinement. This targeted review not only clarifies the current state of research in this field but also offers insights into future research directions, particularly in the development of decision frameworks that move beyond simplistic binary repair actions.

The remainder of this paper is organized as follows. In Sect. 2, we review single-agent DRL approaches for offshore wind maintenance, highlighting key algorithms and their performance in various decision-making settings. Section 3 extends the discussion to multi-agent DRL frameworks, which address scalability by distributing decisions across multiple agents. In Sect. 4, we detail the formulations used to represent the maintenance problem, including Markov decision processes (MDPs), partially observable MDP (POMDP), and hierarchical and graph-based methods. In Sect. 5 we discuss the applications of DRL and the integration of domain-specific knowledge, such as wind farm aerodynamics, weather constraints, logistics, and PHM data, and finally offer a recap of the reviewed models, schematizing their key features in a summary table. Section 6 summarizes the key contributions and performance improvements achieved by the reviewed DRL models, and we also compare simulation-based studies with real-world applications and discuss the integration of nuanced repair types in the models. In Sect. 7 we focus on what we believe is the main gap in current DRL models for maintenance planning, i.e. their reliance on a binary maintain-or-not decision. Instead, we argue that incorporating multiple levels of repair actions is necessary to reflect real-world maintenance scenarios more accurately. Finally, Sect. 8 concludes with insights and directions for future research.

2 Single-agent DRL approaches for offshore wind O&M

Most DRL-based maintenance planners for offshore wind adopt a *single-agent* paradigm, where one agent learns an optimal policy for the entire system (e.g. a wind farm or a single turbine). This structure is suitable when a central decision-maker can coordinate all maintenance actions and information is aggregated at the farm or turbine level.

In DRL, two key functions define the learning objective: the *state action value function* $Q(s,a)$, which estimates the expected cumulative reward of taking action a in state s, and the *state value function* $V(s)$, which measures the expected

reward of being in state s and following the current policy. Different algorithm families approximate and use these functions in distinct ways; thus, among single-agent methods, three main families of DRL algorithms are most frequently applied:

Value-based. Value-based methods, such as the deep Q network (DQN) algorithm (Mnih et al., 2015), learn an approximation of $Q(s,a)$ using a deep neural network and select actions that maximize this value. Stability is achieved through experience replay and a target network. DQN and its variants (double DQN, duelling DQN) are effective for discrete maintenance decisions such as *maintain* versus *not maintain*.

Policy gradient. instead of estimating value functions, policy-gradient methods learn a parameterized policy $\pi_\theta(a|s)$ by adjusting the parameters θ in the direction of the performance gradient. Proximal policy optimization (PPO) (Schulman et al., 2017) is a widely used variant that constrains policy updates within a clipped trust region, improving stability and sample efficiency for large or continuous decision spaces, such as allocating multiple maintenance crews.

Actor critic. Hybrid algorithms, such as actor-critic methods, combine both paradigms by maintaining a policy (the *actor*) and a value estimator (the *critic*). Deep deterministic policy gradient (DDPG) (Lillicrap et al., 2016) and soft actor critic (SAC) (Haarnoja et al., 2018) extend DRL to continuous control using deterministic or stochastic policies with off-policy learning, while asynchronous advantage actor critic (A3C) (Mnih et al., 2016) accelerates training through parallel environment instances.

The flowchart in Fig. 2 illustrates these single-agent DRL approaches: all methods share common initial steps (environment setup, state observation) before diverging by learning logic – DQN (green) selects actions via ϵ-greedy exploration based on Q values, policy-gradient methods (purple) sample actions from a learned distribution and update directly via performance gradients, and actor-critic methods (orange) use a critic network to evaluate actions and guide policy improvement.

The following subsections review how each family has been applied to specific O&M formulations and performance objectives.

2.1 Deep Q networks (DQN)

Value-based methods like DQN are popular for discrete maintenance decisions (e.g. whether to service a component now or later). For instance, Kerkkamp et al. (2022) combined DQN with graph neural networks to take into account asset topology. In their framework, a single agent uses a graph convolutional network (GCN) to group maintenance actions on geographically proximate pipes, yielding more efficient schedules. The DQN+GCN approach produced more reliable networks and higher maintenance grouping compared to a plain DQN and to conventional preventive/corrective policies.

Similarly, Lee et al. (2025) developed a domain-informed DQN ensemble to schedule offshore wind farm maintenance tasks. They formulated maintenance scheduling as an MDP and incorporated wind wake effect models and weather variability into the state. By using convolutional layers to process spatial–temporal features (like turbine–wake interactions), their DQN agent improved power generation by 11.1 % compared to a baseline schedule.

These studies chose DQN for its stability in discrete action spaces and supplemented it with domain-informed neural architectures such as convolutional neural networks (CNNs) or graph neural networks (GCNs) to accelerate learning and capture dependencies. Another value-based approach is the double DQN (DDQN) or duelling DQN to handle large state spaces more stably. Zhang and Si (2020) tested DDQN for large state spaces in multi-component maintenance and showed better performance than simple threshold policies.

However, value-based methods can struggle when the action space or planning horizon grows large or when the state is partially observed (e.g. uncertain component health) (Hausknecht and Stone, 2015a). These challenges have led researchers to explore policy-gradient methods as well.

2.2 Policy-gradient methods

Policy-based DRL algorithms directly optimize a parameterized policy $\pi(a|s)$ and are particularly advantageous in long-horizon, stochastic, or partially observable environments, which are characteristic of offshore maintenance planning. Because policy-gradient updates do not require explicit enumeration of all Q values, these methods handle continuous or large multi-discrete action spaces more naturally than value-based approaches. They also tend to yield smoother and more stable optimization when rewards are sparse or delayed (Schulman et al., 2017), which helps explain the successful use of PPO in discrete maintenance-planning studies such as Pinciroli et al. (2021) and Cheng et al. (2023).

Pinciroli et al. (2021) developed a PPO-based agent to optimize maintenance dispatch for a wind farm with multiple crews. They formulated the problem as a sequential decision process and included prognostic information in the state (predicted RULs of turbines from PHM systems and even forecasted power production for upcoming days). The PPO agent learned a policy that outperformed corrective, scheduled, and threshold-based predictive maintenance benchmarks in profit maximization. Notably, the DRL policy automatically scheduled maintenance during low-power periods and anticipated failures using RUL predictions, something a simple RUL threshold policy could not do. PPO effectively handles con-

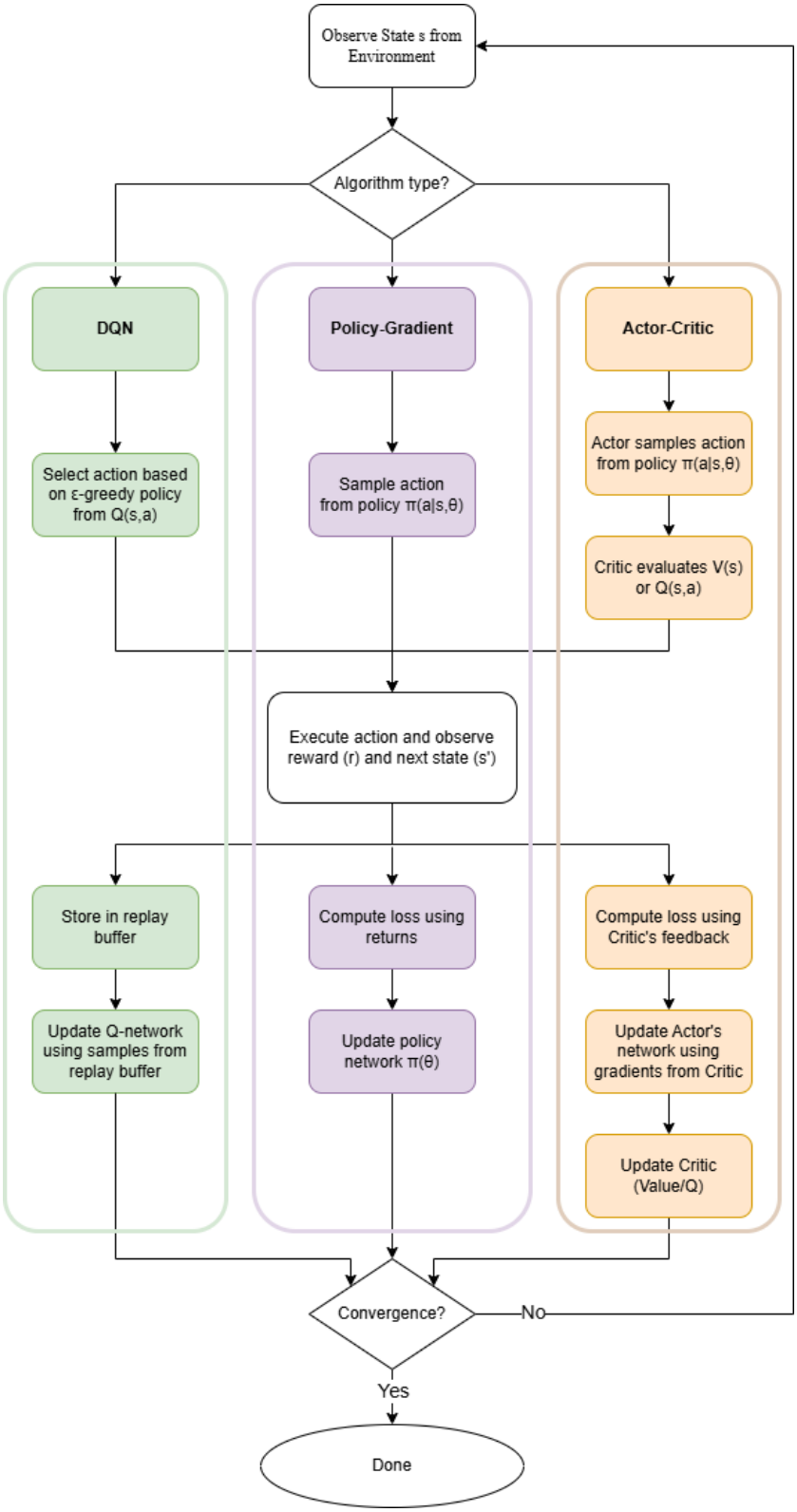

Figure 2. Overview of single-agent deep reinforcement learning algorithm families: DQN (green), policy gradient (purple), and actor critic (orange).

tinuous decision-making under uncertainty and yields better long-horizon rewards than static policies.

In another example, Cheng et al. (2023) applied both DQN and PPO to learn cost-optimal condition-based maintenance for an offshore turbine component. They investigated policies with dynamic inspection intervals and adaptive repair thresholds, formulating the decision as an MDP and comparing DRL algorithms. Both a DQN agent and a PPO agent were able to discover optimal policies under varying conditions, outperforming fixed-interval or fixed-threshold strategies by reducing lifecycle costs, although, in their case, PPO performed better than DQN.

2.3 Actor critic and others (DDPG, SAC, A3C)

In problems with continuous-action decisions (e.g. scheduling exact timing), actor-critic methods like the deep deterministic policy gradient (DDPG) (Lillicrap et al., 2016) and soft actor critic (SAC) become relevant. In a broader manufacturing maintenance context, for example, A3C has been used to optimize resource allocation problems where quick convergence was needed (Mnih et al., 2016).

In the context of wind energy, Zhao and Zhou (2022) used an SAC algorithm to solve a maintenance scheduling problem formulated as an MDP. SAC's ability to handle continuous actions, maximizing the trade-off between reward and entropy, makes it suitable for maintenance problems requiring fine timing control. Their implementation allowed the agent to decide whether to perform maintenance on a turbine in each time period, aiming to minimize long-term cost. The SAC-based scheduler showed improved adaptability to random wind and failure events, outperforming greedy or periodic policies in simulation. Asynchronous advantage actor-critic (A3C) and related algorithms have also been mentioned in the context of maintenance optimization to benefit parallel training. While we are not aware of specific applications of A3C to offshore wind maintenance, the algorithm's ability to run multiple environment instances in parallel can accelerate training for complex simulations.

In principle, A3C/A2C could be applied to wind farm O&M planning to speed up learning across many simulated weather scenarios or failure realizations. However, stability can be an issue, so recent works have gravitated more to PPO for O&M problems.

To provide a clearer overview of how these families compare in terms of strengths, limitations, and typical use cases in O&M decision-making, Table 1 summarizes their key characteristics.

The discussion on single-agent DRL approaches has highlighted how tailored algorithms can effectively optimize maintenance decisions. However, as wind farms scale up and the complexity of interdependent maintenance decisions increases, a shift toward distributed decision-making might become necessary. The following section explores multi-agent DRL frameworks, which distribute the decision-making process across multiple agents, thereby addressing scalability challenges while enabling coordinated maintenance planning.

3 Multi-agent DRL in maintenance planning

As wind farms scale up and the environment increases in size, a multi-agent DRL (MADRL) approach can be considered to distribute the decision-making across multiple agents (Lowe et al., 2017) (e.g. one agent per turbine or per subsystem).

Figure 3 contrasts the single-agent and multi-agent DRL approaches. In the single-agent framework (left), one centralized policy observes the environment and makes all maintenance decisions, whereas in the multi-agent framework (right), each agent receives local observations and collectively coordinates actions.

Multi-agent frameworks address the curse of dimensionality that a single agent faces when managing many components simultaneously (Ogunfowora and Najjaran, 2023). In a cooperative MADRL setting, agents must learn policies that jointly optimize the overall maintenance outcome. A common approach is centralized training with decentralized execution: during learning the agents share information, but in execution each acts independently (Rashid et al., 2018).

A multi-agent perspective is taken by Andriotis and Papakonstantinou (2021), who developed a deep centralized multi-agent actor-critic (DCMAC) algorithm. This algorithm treats each component as an "actor" with individual actions and uses a centralized critic to evaluate the joint outcome. Their method allowed individualized component-level decisions (like a multi-agent system) but maintained a single value function for the overall system. This hybrid approach achieved strong results on high-dimensional maintenance problems, outperforming time-based and condition-based benchmarks. DCMAC can be seen as bridging single- and multi-agent methods: it is centrally trained on the whole state, but action vectors are factorized per component. The success of DCMAC and similar centralized-training approaches again underlines that decomposing the action space among multiple decision-makers is a powerful strategy for scalability.

Another relevant study is Nguyen et al. (2022) (expanded in Do et al., 2024), who optimized maintenance for a 13-component system using a weighted QMIX algorithm. Here, each component is controlled by an agent, and a mixing network combines their action values into a global Q value to enforce cooperation.

The "weighted" QMIX (W-QMIX) variant addresses limitations of the standard QMIX by improving credit assignment to each agent's actions (Liang et al., 2025). By customizing QMIX, they overcame the exponential growth of joint action space and achieved cost-effective policies that significantly reduced total maintenance cost compared to in-

Table 1. Comparison of single-agent deep reinforcement learning algorithm families for offshore wind operations and maintenance.

Algorithm family	Strengths	Limitations
Value based (DQN, DDQN, duelling DQN)	Effective for low-dimensional discrete decisions Sample efficient with experience replay Simple and stable for small action spaces	Less stable in long-horizon problems Sensitive to partial observability Harder to scale to multi-discrete decisions
Policy gradient (PPO)	Stable updates in long-horizon, stochastic settings Robust in partially observable environments Naturally handles multi-discrete decision elements	Less sample efficient than value-based methods Performance sensitive to reward shaping No inherent advantage in simple discrete tasks
Actor critic (A2C/A3C, SAC, DDPG)	Combines policy stability with value-based guidance Efficient for complex states or long dependency chains	Training can be unstable without tuning Off-policy AC methods require careful replay-buffer design Computationally heavier than DQN variants

Figure 3. Comparison of single-agent and multi-agent deep reinforcement learning approaches.

dependent or rule-based policies. In fact, their multi-agent policy outperformed a traditional threshold-based maintenance strategy, yielding 20 % lower cost in the case study. The agents learned to coordinate, essentially performing opportunistic maintenance on other components when one component required service, something that is hard-coded in opportunistic heuristics but emerged naturally via learning. Notably, they simplified the architecture by using a branching neural network (one network outputting multi-component actions).

Figure 4 presents a high-level overview of how local Q values from multiple agents are combined into a single joint Q value through a central mixing network. At the top, global state observations and individual agents' local Q values feed into the network. Within the mixing network, these local estimates are merged. The output at the bottom is a single joint Q value, enabling decentralized agents to learn coordinated policies that account for global objectives.

While most offshore wind DRL studies to date use a single centralized agent, the multi-agent perspective is highly relevant. A wind farm can be seen as a team of turbines (or a team of maintenance crews) that could each be an agent. Multi-agent DRL can explicitly model interactions like shared resources (e.g. a vessel cannot fix two turbines at once) and learn decentralized policies. For example, one could have a DRL agent for each maintenance vessel coordinating via a mixing network to maximize farm availability.

Nevertheless, multi-agent approaches introduce challenges such as non-stationarity during training and coordination complexity. To avoid these issues, for example, Pinciroli et al. (2021) effectively used a single PPO agent to dispatch multiple crews, which could be interpreted as a centralized multi-action policy rather than fully decentralized agents.

An interesting future direction is to combine multi-agent RL with the physical layout of a wind farm, e.g. treating turbines as agents that learn when to request maintenance or crews as agents learning which turbine to service to further scale O&M optimization. In summary, multi-agent DRL is a promising direction that can address scalability and modularity in offshore wind maintenance, ensuring that solutions remain effective as the number of assets grows.

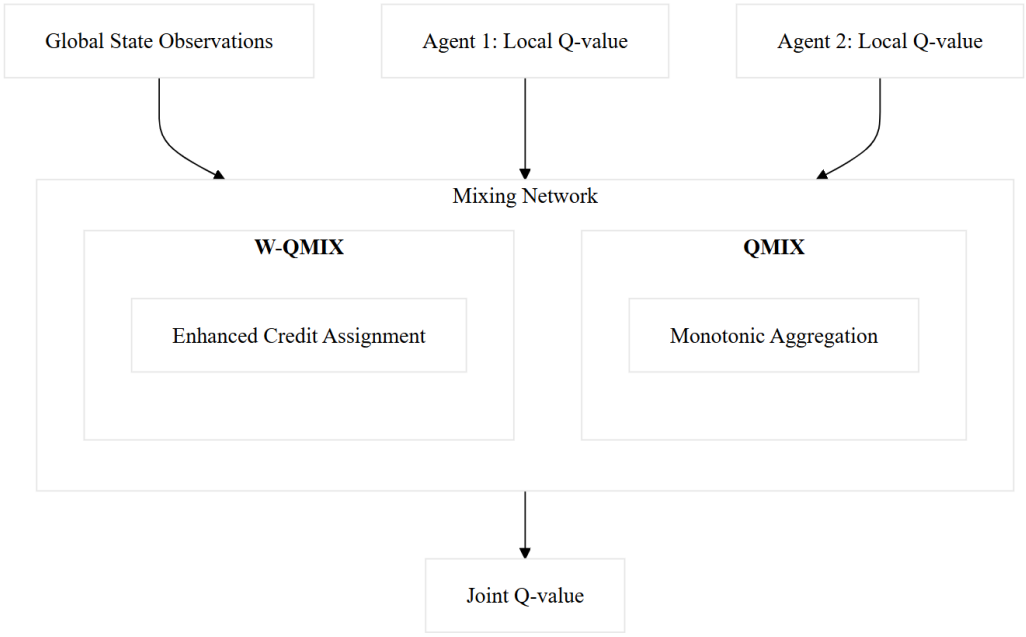

Figure 4. Multi-agent mixing network architecture.

Having examined both single-agent and multi-agent DRL frameworks for maintenance planning, it is clear that the choice of algorithm is intricately linked to how the underlying problem is modelled. The next section focuses on the core methodologies and decision frameworks used, ranging from Markov decision processes (MDPs) and partially observable MDPs (POMDPs) to hierarchical and graph-based representations, which provide the foundation for these DRL approaches. This discussion will clarify how the characteristics of the problem formulation drive the design and performance of the DRL models.

4 Problem formulation

Formulating the maintenance planning problem correctly is a core requirement for DRL. In the literature, we see formulations as MDPs, POMDPs, and even graph-based representations, each chosen to capture the nature of the decision environment.

4.1 Markov decision process (MDP)

Many works assume the system state is fully observable. For example, if one uses direct sensor readings or known component health states as the state, the decision process can be modelled as an MDP.

The reward design typically combines negative maintenance costs and downtime losses into a single scalar reward (or profit) to guide the agent toward cost-optimal decisions, as in Pinciroli et al. (2021).

The CBM decision can also be formulated as an MDP where the state might include the current damage level or time since last inspection as in Cheng et al. (2023). They defined actions such as whether to inspect or repair, with rewards based on cost.

Similarly, the DQN-based scheduling by Lee et al. (2025) uses an MDP where the state includes turbine power outputs (affected by wake and weather) and maintenance statuses, assuming those are known to the agent.

MDP formulations are simpler and allow the use of standard DRL algorithms, but they rely on having a reliable estimator of the system's health state.

4.2 Partially observable MDP (POMDP)

Several offshore wind O&M decision processes are more naturally modelled as POMDPs because the agent typically does not observe the true underlying system state but only noisy and delayed measurements (SCADA/CM signals, inspection outcomes, imperfect weather and access forecasts). In a POMDP, the agent maintains a belief to make decisions under uncertainty (Kaelbling et al., 1998).

Beyond SCADA-based monitoring, a complementary body of literature focuses on non-destructive evaluation (NDE) techniques that can provide inspection- and SHM-driven observations for wind turbine components and structures (e.g. blades, tower, foundations). For instance, Civera and Surace (2022) review non-destructive techniques for wind turbine condition and structural health monitoring over the last 2 decades, covering approaches such as visual inspection, acoustic emission, ultrasonic testing, infrared ther-

mography, radiographic and electromagnetic methods, and oil monitoring, with deployments ranging from human inspections to robotic and UAV-based surveys. Such sensing and inspection modalities can enrich the observation space available to data-driven O&M decision support, but they also highlight why offshore maintenance planning is often partially observable in practice: measurements can be sparse in time (inspection driven), noisy, and operationally constrained by access windows and logistics, motivating POMDP formulations and memory-/belief-based policy representations.

In practical DRL implementations, three families of remedies are commonly used to mitigate partial observability and approximate belief-state reasoning:

History-based state augmentation. A simple approach is to concatenate a fixed window of past observations and actions to the agent input, thereby providing short-term memory of recent dynamics. While straightforward, this can be insufficient when relevant dependencies extend over long horizons (e.g. degradation accumulation, delayed maintenance effects) (Kaelbling et al., 1998).

Recurrent DRL (implicit belief state via memory). Recurrent neural networks (typically LSTMs/GRUs) can compress the observation action history into a hidden state that acts as an implicit belief surrogate. This idea has been adopted in value-based learning (e.g. deep recurrent Q networks, DRQNs) (Hausknecht and Stone, 2015a) and in actor-critic/policy-gradient settings, where recurrent policies are trained end to end to improve performance under partial observability (Williams, 1992; Schulman et al., 2017).

Transformer-based memory (long-range dependencies). When long-range temporal structure matters, transformer architectures offer an alternative to RNNs by using attention mechanisms to retrieve relevant past information. Transformer-based RL agents have demonstrated improved stability and credit assignment in partially observable and long-horizon tasks by enabling flexible access to historical context (Parisotto et al., 2020). This is conceptually aligned with offshore maintenance planning, where optimal actions may depend on events far in the past (e.g. prior repairs, earlier inspections), although transformers may require careful regularization and substantial training data.

Beyond implicit memory, explicit belief-state estimation can also be pursued by combining filtering with RL, for instance using Bayesian filters or particle filters when a tractable transition/observation model is available or by learning latent-state models that infer hidden degradation dynamics from observations before or during policy learning (Kaelbling et al., 1998; Igl et al., 2018). Overall, expanding offshore wind DRL formulations from MDP to POMDP settings primarily affects the policy representation: memory-augmented policies (recurrent or transformer based) and/or explicit belief estimators provide practical mechanisms to cope with uncertainty in health information, accessibility, and delayed maintenance outcomes.

A first example of such a formulation for an O&M planning problem is Kerkkamp et al. (2022), who formulate maintenance planning as a POMDP on a graph, where the underlying deterioration states are partially observed through inspections.

Similarly, Lee and Mitici (2023) treat their predictive maintenance problem for aircraft components as a POMDP, using a CNN to process raw sensor data into an observation for the DRL agent.

The choice of POMDP acknowledges that maintenance decisions must be made with imperfect information, and DRL agents in this setting are trained to be more robust to uncertainty (e.g. scheduling maintenance a bit earlier to hedge against uncertain failure times). Methodologically, solving a POMDP with DRL often means using belief state features or statistical features of uncertainty (such as predicted failure probability) in the state.

4.3 Graph- and network-based representations

Some complex systems benefit from graph-based formulations. The study by Kerkkamp et al. (2022) is an example of where the asset network structure (a sewer network in their case) is encoded as a graph, and a GCN is used to inform the DRL agent.

While their domain was not wind, one can imagine an offshore wind farm graph where nodes are turbine components and edges could represent spatial proximity or electrical/-functional dependencies. This approach could let the agent learn policies that consider component interactions (e.g. if neighbouring turbines' maintenance can be combined). In their framework, the graph-based state and GCN embedding encouraged the agent to group maintenance geographically, improving efficiency.

4.4 Hierarchical and interpretable models

To tackle the black-box nature of DRL, Abbas et al. (2024) introduced a hierarchical DRL framework for turbofan engine maintenance that combines an input–output hidden Markov model (IOHMM) with a DQN-like agent. The high-level IOHMM module interprets sensor data to detect the likely fault mode or degradation state (providing a human-understandable diagnostic), while the low-level DRL module learns the optimal replacement or repair policy given that inferred state. This two-level approach achieved performance on par with end-to-end DRL but with the added benefit that decisions could be traced to identifiable health-state estimates.

In safety-critical domains like aerospace or offshore energy, such interpretability is valuable for gaining trust in the AI's recommendations. Moreover, by narrowing the policy

search to focus on critical decisions (informed by the HMM), the agent avoids spurious actions in sparse failure domains.

Although turbofan engines differ from wind turbines, the concept is relevant: they use a probabilistic model to identify health states and provide an interpretable layer, and the DRL agent makes maintenance decisions at a higher level. This yields a more transparent policy, which could be valuable in offshore wind where operators demand an understanding of the AI's decisions (Adadi and Berrada, 2018). Such hierarchical or hybrid frameworks (combining physics-based or expert models with learning) are a way to inject domain knowledge directly into the DRL algorithm, improving learning speed and trustworthiness.

Figure 5 illustrates different frameworks commonly used to represent state information in deep reinforcement learning (DRL) models for maintenance decision-making. Starting from an initial state S, the flowchart shows four distinct ways the state can be represented or processed before making a decision. The MDP formulation (blue) assumes full access to the true system state. The POMDP approach (yellow) highlights the partial observability of real-world systems, requiring the agent to infer underlying states from limited observations $O(S)$ of the true state. The graph-based method (green) structures state information using node and edge relationships to capture spatial or topological dependencies, allowing the agent to consider how interconnected assets affect one another's maintenance needs. Finally, the hierarchical approach (red) incorporates domain-specific insights, from either expert knowledge or physics-based models. After selecting an action A based on the chosen representation, the system receives a reward R and transitions to a new state S', looping back to begin a new decision-making cycle.

4.5 Decision frameworks and objectives

Across these formulations, the decision-making frameworks can vary in objective. Some agents aim to maximize availability or energy production (profit) (Pinciroli et al., 2021; Lee et al., 2025). Others aim to minimize total cost (including repair costs, downtime costs, and a possibly penalty for using resources) (Kerkkamp et al., 2022).

These are effectively two sides of the same coin, since maximizing uptime or output will implicitly minimize downtime losses. The reward function should, in fact, include terms for lost revenue when a turbine is down, crew/vessel dispatch costs, spare part costs, and maybe even penalties for equipment degradation.

For example, Pinciroli et al. (2021) designed a reward equal to the short-term profit (energy revenue minus O&M costs) at each decision step, so the PPO agent's cumulative reward corresponds to total profit. Cheng et al. (2023) focused on cost rates, giving negative rewards for inspection or repair costs and for failures, to encourage the agent to find the policy with minimum average cost.

In multi-agent settings, a global reward is often used for full cooperation (as in QMIX, where agents maximize a shared cost-saving metric).

Some works also impose constraint handling in the formulation. For instance, maintenance actions might be invalid under certain weather conditions; a realistic environment simulator will simply not allow those actions (or will assign a large negative reward if attempted). Thus, ensuring decisions are feasible (e.g. not sending a crew when waves are too high) can be done by action masking or via constraint penalty in the reward. Recent research even explores incorporating optimization constraints into neural network design (e.g. using attention masks to enforce constraints) as in Kazemian et al. (2024), though this is still emerging in O&M applications.

In summary, the methodologies range from straightforward MDP models with fully observable states to sophisticated POMDP and graph-based models that embrace the complexities of offshore wind maintenance. The trend is toward more realistic problem formulations, acknowledging partial observability, incorporating spatial and logistical structure, and aligning the reward with business metrics (cost or profit). The following section provides a structured summary of these methodologies based on their application and the domain-specific knowledge considered, finally highlighting key aspects such as agent design, algorithm choices, and problem formulations in a comparative table. This summary aims to offer a clear and concise reference for understanding the variations in DRL-based maintenance planning approaches and their defining characteristics.

5 Applications of DRL approaches for offshore wind O&M

A fundamental aspect of training and deploying DRL agents for maintenance planning is the integration of domain-specific knowledge. The following subsections explore how different forms of domain knowledge can enhance decision-making: (i) *wind farm aerodynamic* interactions that affect energy yield; (ii) *weather and sea state* constraints that determine accessibility; (iii) *logistics* and crew-related considerations; (iv) *prognostic and health management data* for predictive maintenance; and, finally, (v) *economic factors* such as market prices and budget limits.

5.1 Wind farm aerodynamics

In a wind farm, turbines cast aerodynamic wakes that reduce the output of downwind turbines (Vermeer et al., 2003). In Lee et al. (2025), the authors considered this by making their DRL agent wake-aware. They fed a wake interaction model into the state and used an ensemble of DQN models to capture different wake scenarios. By incorporating multiple wake models, their agent learned maintenance decisions that minimize farm power loss due to wakes, leading to higher overall energy production.

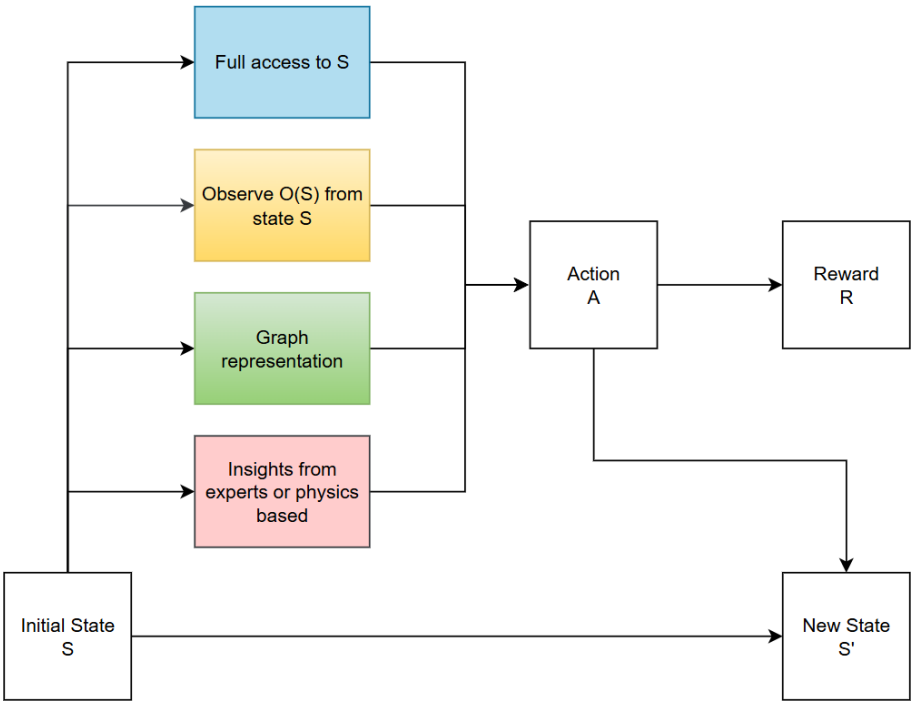

Figure 5. Overview of problem formulations in deep reinforcement learning models: Markov decision process (blue), partially observable Markov decision process (yellow), graph-/network-based formulation (green), and hierarchical approaches (pink).

This domain knowledge is particularly important for tightly spaced wind farms where wake losses are significant; a purely self-learned agent without wake inputs might need enormous training experience to "discover" such effects, whereas providing a wake model upfront accelerates learning.

The result is a policy that schedules maintenance such that either the waking effect is minimal (e.g. performing maintenance when winds are low or from directions that do not affect other turbines) or multiple affected turbines are maintained together to collapse the wake loss into a single period.

5.2 Weather and sea state

Offshore maintenance is heavily weather dependent, as high waves, strong winds, or storms can prevent crew transfers and repairs (Jenkins et al., 2021). Some DRL studies explicitly include weather in the environment.

Chatterjee and Dethlefs (2021) demonstrated a DRL approach for planning vessel transfers to turbines, integrating real SCADA data and weather conditions. Their agent could prioritize critical repairs and navigate the stochastic availability of weather windows, something traditional scheduling struggles with (Borsotti et al., 2024). By training on historical weather patterns, the DRL policy learns, for example, to take advantage of a calm sea state to perform a repair even if it is slightly early because waiting might mean a long weather delay.

Similarly, Pinciroli's state included predicted power (which indirectly reflects wind forecast) to help the agent plan around periods of low wind (often correlated with calmer weather) (Pinciroli et al., 2021; Ogunfowora and Najjaran, 2023).

In practice, one can input wave height forecasts or wind speed forecasts into the DRL state; the agent will then learn not to "choose" an action that requires travel during bad weather. Domain knowledge here ensures feasibility and robustness: the agent that knows about weather will inherently develop a maintenance schedule that aligns with seasonal weather patterns, reducing cancellations and idle times.

5.3 Logistics

Offshore O&M involves vessels, helicopters, crews, and spare parts – logistical aspects that greatly affect cost. DRL models have started to include these. In the multi-crew PPO model by Pinciroli et al. (2021), the state and action were designed to capture crew positions and availability. The environment simulation accounted for travel times to turbines and repair durations. This domain realism meant the learned policy actually coordinates crew movements: e.g. sending Crew A to a turbine that will finish repair soon, while Crew B waits at the depot until a large failure occurs. By encoding travel time and multiple crews, the DRL agent learned to avoid wasted trips and to keep crews busy, emulating optimal routing and scheduling decisions.

In Kerkkamp et al. (2022), logistics is addressed in a spatial sense by using a GCN to group geographically close maintenance. In an offshore wind context, that could translate to handling nearby turbines in one outing to minimize transit.

Even without an explicit graph, a DRL agent can learn from cost feedback that doing maintenance on neighbouring turbines back to back saves the transit cost of multiple separate trips. Future research should incorporate more detailed logistics, such as vessel capacity, fuel cost, and inventory of spare parts, into the DRL state/reward. The benefit of doing so is that the learned policy becomes a holistic O&M schedule that respects not just failure risks but also supply chain and labour constraints.

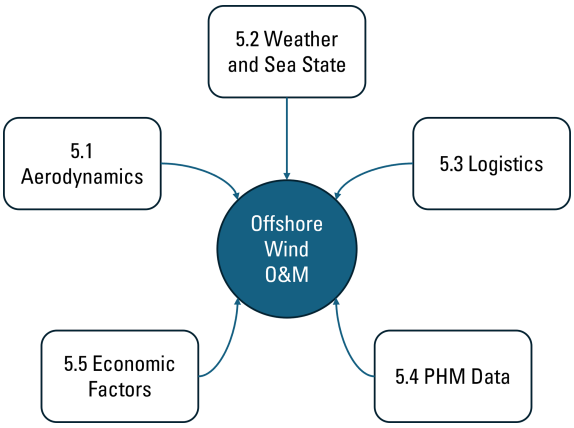

Figure 6. Domain-specific knowledge for offshore wind.

5.4 Prognostics and health management (PHM) data

Almost all DRL approaches in this area use PHM outputs (like condition monitoring and RUL predictions), which is essential domain knowledge for predictive maintenance. The difference is in how they incorporate it. For instance, one could use a discretized health state (e.g. "good", "degraded", "critical") as part of the state space, which is easier for an agent to handle than raw sensor readings. The approach of Lee and Mitici (2023) of using a CNN on sensor data before the DRL agent is another way, essentially letting a deep learning model extract relevant features (e.g. vibration patterns) which portray the health knowledge.

For different applications, Abbas et al. (2024) and Abbas (2024) integrate an input–output hidden Markov model to classify health states of turbofan engines and feed this into the DRL agent.

By combining PHM and DRL, these frameworks close the loop from condition monitoring to decision-making. The advantage is clear: the better the agent's awareness of actual component condition (even if inferred), the better it can time maintenance. PHM can also be embedded in the model by shaping the reward as well; e.g. a large penalty for a failure effectively encodes the idea that "a catastrophic gearbox failure is very bad", which the agent must learn.

PHM-driven rewards can also be used, such as giving a small penalty for running a component in a highly degraded state (reflecting higher wear or risk).

5.5 Economic factors

Domain knowledge also includes economic factors like energy price and maintenance cost structure. Some works incorporate dynamic electricity prices or contractual penalties into the reward. For instance, if energy price forecasts are available, a DRL agent could decide to do maintenance when energy price (hence lost revenue) is low.

While not explicitly seen in the reviewed models, attention-based approaches are starting to include market price as part of the input to O&M scheduling models (Kazemian et al., 2024).

In offshore wind, including such factors could align maintenance not just with technical needs but also with business cycles (e.g. perform maintenance during low-demand seasons or when subsidy/price is low). Another economic constraint is the budget or resource limit. DRL can consider these by capping certain rewards or through state variables (like remaining budget).

Integrating domain-specific information can make DRL formulations more realistic and better aligned with the physical and operational characteristics of offshore wind systems. Figure 6 summarizes the types of knowledge commonly incorporated in existing studies, including wake interactions, weather accessibility, logistics constraints, degradation physics, and PHM-derived indicators. The role of these elements is to provide structure that may help the agent focus on features that are relevant to decision-making; nevertheless, it is worth noting that increased model complexity, additional training effort, or conflicting inputs may offset potential gains, and over-specifying the environment can make optimization more difficult rather than easier. Thus, domain knowledge can support DRL when carefully selected and validated, but its contribution depends on the quality, relevance, and reliability of the information provided.

All the studies reviewed have, in one way or another, blended domain knowledge into their DRL approach: Lee et al. (2025) added wake/convective layers, Chatterjee added weather data, Kerkkamp added graph relations, Pinciroli added RUL and power forecasting, Abbas added HMM interpretable layers, etc. This synergy of domain expertise and reinforcement learning is key to developing trustworthy, high-performance maintenance policies that industry can adopt.

To compare the DRL-based maintenance planning approaches reviewed, Table 2 highlights their key features, such as agents, algorithms, and problem formulations. All the mentioned approaches aim to optimize maintenance scheduling but differ in the algorithms and formulations used. The

column *agent* refers to whether one policy controls the whole system or multiple coordinating policies exist, *algorithm* shows which training method was used, *problem formulation* indicates how the maintenance decision problem is modelled for the agent, and *domain-specific knowledge* indicates which categories of offshore-wind-specific knowledge are explicitly integrated in each study (e.g. wake/aerodynamic effects, weather, logistics, and PHM information).

In Fig. 7, we provide a comparative view of the distribution of key features across the reviewed literature. The pie chart on the left shows the proportion of single- versus multi-agent frameworks. The central pie chart outlines the distribution of algorithms, such as DQN, PPO, SAC, and QMIX variants, showing that DQN remains the most commonly used DRL algorithm despite the explosion of the action space, which grows with the number of components or maintenance tasks, limiting its applicability in larger-scale problems. Finally, the pie chart on the right highlights the prevalence of MDP- and POMDP-based modelling.

With a clear understanding of the diverse methodologies used to model offshore wind maintenance planning, we now turn to the practical impact of these approaches. The following section summarizes the key contributions and performance improvements achieved through the application of DRL, illustrating how these methodological choices translate into tangible benefits such as cost reduction, improved reliability, and enhanced operational efficiency.

6 Discussion

The application of DRL to offshore wind O&M has shown clear advantages over conventional maintenance strategies, achieving lower costs, higher availability, and more adaptive planning. The reviewed studies highlight five recurring benefits: (i) *cost and downtime reduction*, as agents learn to time interventions "just in time" before failure; (ii) *enhanced predictive maintenance*, through integration of remaining useful life (RUL) data and operational context; (iii) *opportunistic maintenance*, by grouping actions and exploiting favourable conditions; (iv) *improved reliability and safety*, via proactive scheduling and embedded risk constraints; and (v) *computational scalability*, enabling optimization of large, stochastic systems.

Finally, we focus on the remaining gaps regarding the use of real-case scenarios for testing the models and the inclusion of multi-level repair that should reflect real operational practices.

The following subsections discuss each benefit in detail, illustrating how DRL contributes to more cost-efficient and resilient offshore wind maintenance planning.

6.1 Cost and downtime reduction

A primary goal is lowering maintenance costs and turbine downtime compared to baseline strategies (reactive or scheduled). DRL agents have demonstrated the ability to significantly reduce unplanned failures and associated costs.

For example, Cheng et al. (2023) report that their PPO-based adaptive inspection and repair policy (DIAR) achieves an expected life-cycle cost of EUR 2.32×10^4 compared to EUR 3.01×10^4 for the best uniform-interval, fixed-threshold strategy (UIFR), i.e. a 23 % reduction in their case study.

In Lee and Mitici (2023), the DRL-based policy reduced unplanned downtime and maintenance cost in a multi-component system by 95.6 % versus conventional periodic maintenance.

These improvements come from the agent's ability to optimize the timing of maintenance: servicing components "just in time" before failure but not too early to waste life. DRL policies effectively find this sweet spot by continuous learning and adjustment.

6.2 Predictive (PHM-based) strategies

Many wind operators use condition-based triggers (like RUL thresholds from prognostics) to plan maintenance. DRL can enhance these predictive strategies by adding dynamic decision-making. As noted in Pinciroli et al. (2021), the DRL policy outperformed a pure RUL threshold policy by considering not only the component health but also external factors such as power demand and crew availability. In other words, whereas a standard predictive maintenance strategy says "replace component X when its RUL $< Y$ days", a DRL agent might learn "replace component X when RUL $< Y$ days *and* a maintenance team is idle *and* a low-wind period is coming up", thereby minimizing impact. This kind of contextual decision-making led to higher reward (profit) in their experiments. We see a similar theme in Lee et al. (2025): their DQN agent, augmented with wake effect knowledge, boosted energy production beyond what a wake-unaware strategy achieved. By learning the true optimal policy through trial and error, DRL approaches can exceed the performance of both corrective maintenance (which incurs high downtime) and simple predictive rules (which might be myopic or inflexible).

6.3 Opportunistic maintenance

DRL has shown strength in exploiting opportunistic maintenance opportunities that humans or simple policies might miss. For instance, in multi-component scenarios, an agent can coordinate maintenance on multiple turbines in one go to avoid repeated downtime. The PPO agent in Pinciroli's work learned to wait for low-wind output days to schedule maintenance, which is an opportunistic behaviour yielding higher rewards. Huang et al. (2020) observed that, even when not explicitly programmed, DRL agents naturally learn to group multiple repairs together, thereby reducing repetitive downtime and sharing high logistics costs over several interventions. Similar findings can be found in Dong et al. (2021)

Table 2. Comparison of deep reinforcement learning approaches for offshore wind farm maintenance.

Reference	Reported gain(s)	Agent		Algorithm						Problem formulation				Domain-specific knowledge			
		Single	Multi	DQN	PPO	SAC	QMIX	W-QMIX	DCMAC	MDP	PO-MDP	Graph	Hierarchical	Aerodynamics	Weather	Logistics	PHM
Lee et al. (2025)	+11.1 % power generation vs. baseline schedule	✓		✓						✓				✓			
Abbas et al. (2024)	NA	✓		✓									✓				✓
Do et al. (2024)	∼ 20 % more cost-effective vs. threshold-based maintenance (total maintenance cost, case study)		✓				✓	✓		✓							
Lee and Mitici (2023)	95.6 % reduction in unplanned downtime and maintenance cost vs. periodic maintenance	✓									✓						✓
Cheng et al. (2023)	23 % life-cycle cost reduction: EUR 2.32×10^4 vs. EUR 3.01×10^4 (DIAR vs. UIFR)	✓		✓	✓					✓							
Zhao and Zhou (2022)	NA	✓				✓				✓							
Kerkkamp et al. (2022)	NA	✓		✓							✓	✓				✓	
Nguyen et al. (2022)	∼ 20 % more cost-effective vs. threshold-based maintenance (total maintenance cost, case study)		✓				✓	✓		✓							
Pinciroli et al. (2021)	NA	✓			✓					✓					✓	✓	
Chatterjee and Dethlefs (2021)	NA	✓		✓						✓					✓		
Andriotis and Papakonstantinou (2021)	NA		✓						✓		✓						

NA means not available.

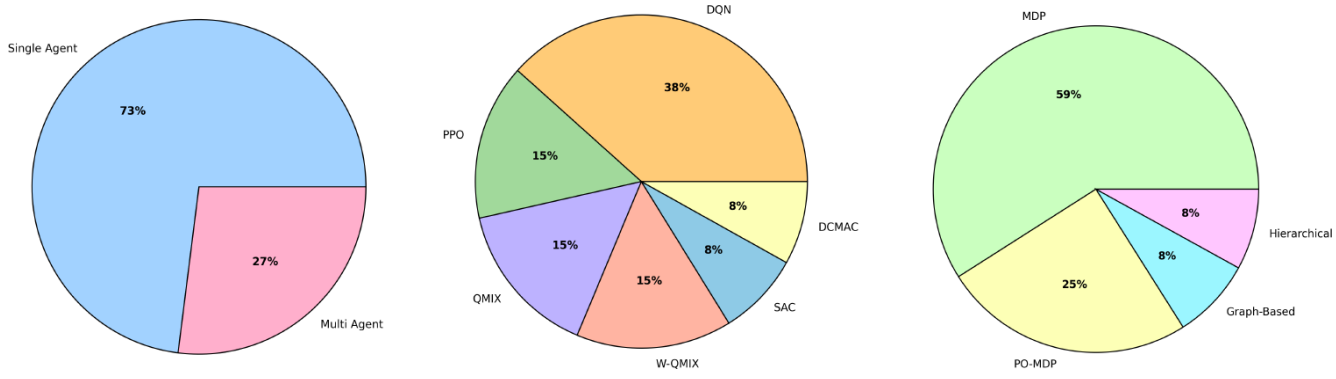

Figure 7. Comparative view of single vs. multi-agent methods, main deep reinforcement learning algorithms, and problem formulations in the reviewed literature.

and Valet et al. (2022). Do et al. (2024) demonstrated this with their multi-agent approach, where the learned policy effectively grouped maintenance tasks to save on shared downtime costs, beating a policy that treats components independently. Even in single-agent setups, agents learn to use times of low production or existing outages to perform additional repairs. In the context of offshore wind, for example, an agent might schedule a minor repair on one turbine when a vessel is already en route for a major repair on a neighbouring turbine, effectively reducing additional transit costs.

6.4 Reliability and safety

An important point is that DRL policies can improve reliability metrics (e.g. mean time between failures, availability) by preventing failures proactively. Lee and Mitici (2023) noted fewer failures in their DRL-maintained system than a conventional approach. Additionally, DRL can incorporate safety constraints (like not allowing maintenance deferral beyond a limit) via rewards or state features. The hierarchical approach by Abbas (2024) is aimed at safety-critical maintenance. By integrating an interpretable model, they ensure the DRL decisions for turbofan engines remain within safe bounds and can be understood by engineers. In offshore wind, ensuring that a DRL policy does not inadvertently run turbines to catastrophic failures is very important. Studies so far show that with proper reward design (heavily penalizing failures), the agents naturally learn to avoid risky deferrals.

6.5 Computational feasibility

Another finding across the literature is that DRL can handle high-dimensional problems that were previously intractable by brute-force optimization. Maintenance scheduling for a wind farm with many turbines, each with several components, is a huge combinatorial problem over a long horizon. Some studies also accelerate learning by transfer or imitation. For example, Pinciroli et al. (2021) initialized their PPO agent via imitation learning from a heuristic policy to shorten training time. This hybrid approach marries human insight with AI optimization. The result is a practical decision-support tool that can quickly recommend which turbine to maintain and when, given the current observations.

DRL, with its experience-driven learning, provides feasible solutions. Nevertheless, training can be computationally intensive (e.g. WQMIX took 12 h in the 13-component case, Nguyen et al., 2022), although once trained, the policy can execute decisions in real-time.

Model-based or brute-force methods struggle to consider all contingencies and long-term effects (Chen et al., 2024).

The consensus of recent work is that DRL-based maintenance planning consistently outperforms static strategies in simulation, often by a wide margin in cost savings or uptime. These improvements stem from DRL's ability to learn optimal scheduling under uncertainty, adapt to varying conditions (weather, load demands), and coordinate multiple decisions in a way that human-designed rules cannot easily mimic.

While the performance gains of DRL-based maintenance strategies are evident, their success is further amplified by the integration of domain-specific knowledge. The next section focuses on how incorporating elements such as wind farm aerodynamics, weather constraints, logistical considerations, and PHM data not only enriches the state representations but also guides the learning process toward more realistic and reliable maintenance policies.

6.6 Simulation-based studies vs. real-world applications

Simulation-based research has been the predominant way to develop and evaluate DRL for wind farm maintenance. All the studies reviewed above rely on simulated environments, typically a stochastic model of turbine degradation and failure, combined with models of maintenance actions (costs, durations, effects) and often weather or logistics simulators.

For instance, Lee et al. (2025) used a simulation of wind dynamics and turbine wakes to train their DQN ensemble; Pinciroli et al. (2021) built a custom simulator for turbine failures and crew assignments to test PPO. Simulation is essential because it provides a safe and flexible sandbox to train DRL agents (which often require millions of decision steps for convergence). Researchers can speed up or repeat scenarios (like many years of operation) to expose the agent to rare events, something impossible to do quickly in the real world.

Key insights from simulation studies include the significant performance gains of DRL policies over traditional maintenance strategies. Many papers report that their learned policies yield lower cost or higher availability than periodic (time-based) or reactive (run-to-failure) maintenance. For example, the agent of Pinciroli et al. (2021) outperformed corrective and age-based schedules, with fewer failures and lower cost, mirroring the results of Andriotis and Papakonstantinou (2021) against periodic policies. Such results, consistently observed in simulation, build a compelling case for applying DRL in practice.

That said, real-world applications of DRL for offshore wind O&M remain at an early stage, and published demonstrations typically rely on simplified physical and operational assumptions. A recent example is the work by Lee et al. (2025), developed in collaboration with an industry research institute (KEPCO). Their DQN-based scheduling framework illustrates how DRL can, in principle, coordinate maintenance actions and exploit wake interactions to improve farm-level energy production. However, their study relies on the Jensen kinematic wake model, which, while computationally efficient, provides only coarse accuracy in the near-wake region and may misrepresent wake recovery dynamics. Consequently, the reported power-gain improvements (i.e. an 11.1 % increase over their baseline scheduling strategy) should be interpreted as gains within the constraints of a simplified simulation environment (Lee et al., 2025). Additional assumptions, such as fixed maintenance durations, deterministic task lists, and the absence of vessel or access constraints, further limit the generalizability of the results.

These studies therefore highlight both the potential and the current limitations of DRL for maintenance planning: DRL can identify useful scheduling patterns in controlled testbeds, but its effectiveness in operational settings remains dependent on the fidelity of the underlying simulator. More comprehensive validation using higher-fidelity wake models, realistic metocean variability, and operational logistics will be

essential to assess the practical applicability of DRL-based maintenance planning.

While the deployment of a trained agent in a live wind farm has not been reported yet, the involvement of an industrial stakeholder and the real-world fidelity of the simulation (including measured wake effects) indicate a step toward actual adoption.

In fields like aviation and manufacturing, some DRL-based maintenance planners have been tested on real data if not deployed directly. For instance, the turbofan case by Abbas et al. (2024) used NASA engine datasets to train and validate their approach. This kind of validation on real-world data builds confidence that the policies will translate from simulation to reality. Moreover, ongoing advances in digital twins for wind farms (Zhang et al., 2024) (high-fidelity replicas of turbines and operations) may enable DRL agents to be trained or at least fine-tuned in an environment that closely mimics reality.

6.7 Multi-level repairs

Traditional maintenance models in offshore wind operations typically reduce decisions to a simple binary choice, either repair or not. However, real-world observations show that turbine downtime arises from a spectrum of failures. For instance, industry data reveal that roughly 70 % of downtime is caused by major repairs, about 17 % by minor repairs, and the remainder due to simple resets (Carroll et al., 2017).

This suggests that maintenance actions in the field are inherently multi-level, ranging from minor fixes that temporarily restore performance to major overhauls or full component replacements.

To make the notion of *multi-level* maintenance actions more concrete in a wind context, Aafif et al. (2022) study preventive maintenance for a wind turbine gearbox under temperature-based condition monitoring. They compare (i) a commonly adopted industrial strategy in which, whenever a threshold temperature is exceeded, the turbine is temporarily derated while the gearbox is cooled, and, after such events become frequent enough, the gearbox is ultimately renewed (replacement or overhaul) against (ii) a multi-level strategy in which each threshold exceedance triggers an imperfect preventive maintenance action that partially restores the gearbox condition by reducing its failure rate to a value between the current rate and that of a new gearbox, with renewal enforced only after N imperfect PM actions. Their numerical comparison shows that neither "renewal only" nor "imperfect PM then renewal" dominates universally: the more economical strategy depends on gearbox reliability and on the relative magnitude of production loss, cooling, preventive maintenance, and renewal logistics costs. This case illustrates why binary "repair/replace" abstractions can be limiting for offshore wind O&M: introducing intermediate action levels (e.g. partial restoration actions before renewal) enables the policy to express realistic trade-offs between short-term production impacts and long-term degradation management.

While incorporating these multi-level actions enriches the state action representation, it also exposes a significant gap in current DRL models for offshore wind O&M. Most existing studies focus on binary decision frameworks.

The lack of integration of multi-level maintenance actions is discussed in greater detail in the next section by examining how DRL-based planning models can incorporate multi-level maintenance strategies, allowing agents to choose among different maintenance tasks. Furthermore, we explore the implications of this refined modelling for cost, reliability, and overall operational efficiency.

7 Future directions

Traditional maintenance models in offshore wind often reduce decisions to a binary choice (e.g. to repair or not). In a DRL formulation, the maintenance planning problem can be recast as a Markov decision process where the state includes asset features such as component condition, age windows, and weather windows, while the action space is expanded beyond a simple "maintain or not" decision.

Instead, this enables the agent to choose among multiple repair options. Such an approach would allow the agent to balance cost and reliability by, for example, deploying a less expensive interim repair when system health is marginally degraded or committing to a full repair only when necessary.

Incorporating minor repairs as a viable action can prevent small degradations from escalating into catastrophic failures. DRL agents trained on multi-level maintenance tasks can learn to execute low-level fixes when early signs of degradation appear, thereby extending component life and improving overall system reliability. Wei et al. (2019), for example, demonstrated that a DRL-based policy for structural maintenance maintained high reliability by optimally balancing minor and major interventions across numerous components.

Potential avenues for enabling a DRL agent to consider multi-action maintenance planning include the following:

Expanded action spaces. Researchers have begun explicitly modelling a range of actions, such as "do nothing", "perform minor repair", "conduct major repair", or "replace component". For example, Zhang et al. (2023) formulate an infinite-horizon DRL maintenance problem where a component's health can be partially recovered through an imperfect repair or fully restored via corrective maintenance. This expanded action space helps the agent learn which level of intervention yields the optimal long-term cost and reliability trade-off.

Parameterized and hybrid actions. To manage the complexity arising from multiple discrete repair choices combined with continuous variables (e.g. the timing of intervention), advanced approaches employ param-

eterized action spaces. In one instance, a parameterized PPO algorithm was developed to handle mixed discrete–continuous decisions, effectively allowing the agent to adjust both the type of repair and the timing simultaneously (Zhang et al., 2023). This structured action space helps maintain convergence stability while exploring complex repair policies.

Multi-agent decomposition. Another promising strategy is to decompose the large-scale maintenance problem by modelling each turbine or even each component as an individual RL agent within a cooperative framework. For example, Su et al. (2022) address the explosion of the action space in multi-level preventive maintenance by treating each machine as an independent agent that coordinates with others. Such a decomposition not only alleviates scalability issues but also enables the agents to learn localized strategies that can later be integrated for holistic wind farm maintenance.

Despite the promise of these approaches, challenges remain in training and deployment. The high dimensionality of both state and action spaces, especially in a wind farm with hundreds of turbines, can lead to a combinatorial explosion of decision possibilities. Techniques such as multi-agent reinforcement learning are being explored to address these scalability concerns. Cross-industry insights from manufacturing, aerospace, and civil infrastructure further suggest that lessons learned in one domain (e.g. mission-aware planning or adaptive grouping strategies) can be effectively translated to offshore wind maintenance planning.

In summary, by moving beyond binary choices, these models could capture the inherent complexity of real-world O&M practices and facilitate the development of policies that more accurately balance short-term fixes with long-term reliability. This advancement would represent a significant step toward more realistic and effective DRL-based maintenance planning in offshore wind operations.

The concluding section synthesizes the insights gathered from the reviewed models, underscoring the strengths, current limitations, and future directions for integrating DRL into operational decision-making frameworks that could revolutionize maintenance planning in the renewable energy sector.

8 Conclusions

The literature demonstrates that DRL is a promising approach for offshore wind farm maintenance planning. By learning from interactions in a simulated environment, DRL agents can devise maintenance policies that outperform traditional corrective, time-based, and other predictive strategies on key metrics like cost, downtime, and energy production. Both single-agent and multi-agent frameworks have been explored: single-agent DRL has succeeded in optimizing complex maintenance schedules by considering long-term consequences, while multi-agent DRL offers a path to scaling these solutions to larger systems by decentralizing decisions. Moreover, the most successful studies embed domain-specific knowledge, from wake physics and weather patterns to prognostic models, into the DRL process, creating hybrid solutions that learn efficiently and behave realistically.

The reviewed works highlight several advantages of DRL in offshore wind O&M: adaptive scheduling that responds to the actual condition of turbines, opportunistic maintenance that smartly times actions to minimize impact on operations, and the ability to handle the high dimensionality of scheduling problems that defy mathematical or brute-force optimization. For example, DRL agents have learned to schedule maintenance during low-wind periods (Ogunfowora and Najjaran, 2023), to cluster repairs and save vessel trips (Kerkkamp et al., 2022), and to prevent failures by reacting to prognostic alarms better than fixed rules. These capabilities translate into quantifiable gains: double-digit percentage improvements in cost savings or energy output in case studies are common (Lee et al., 2025; Nguyen et al., 2022).

However, challenges remain to be solved before DRL becomes commonplace in live offshore wind operations. Safety and trust are critical aspects: operators need assurance that an AI agent would not recommend catastrophic decisions. This is why interpretability (as in the hierarchical HMM approach) and extensive testing are crucial. Computational efficiency is also a concern: multi-agent or long-horizon DRL can be computationally intensive in the training phase, though improvements in algorithms and hardware mitigate this. Additionally, integration with existing maintenance management systems requires user-friendly interfaces and perhaps human-in-the-loop designs (where human planners can review or override AI suggestions).

Moreover, a critical limitation in current DRL applications is the overly simplistic treatment of repair actions. Most methods consider only one kind of repair, typically reducing the decision to whether to replace a component. In practice, maintenance is a multifaceted process that often involves choosing among various repair strategies, each with distinct implications for system performance and cost. For instance, an optimal policy should not only determine the optimal timing of an intervention but also decide which specific repair tasks to undertake based on the current condition of components. Addressing this gap requires developing agents capable of discerning a richer set of actions that reflect the complexities of real-world maintenance tasks.

For practitioners and researchers, these findings support moving toward DRL-based decision support tools for wind farm O&M. As offshore wind farms continue to grow and data from operations accumulate, DRL approaches aim to become integral in optimizing maintenance planning, ultimately lowering the cost of renewable energy and improving the reliability of wind power generation.

Despite efforts to follow a transparent and reproducible literature selection protocol as defined in Sect. 1, several threats to validity remain. First, *publication bias* may affect the evidence base: studies reporting positive performance gains of DRL approaches are more likely to be published than negative or inconclusive results, and industrial deployments are often underreported due to confidentiality. As a consequence, the reviewed corpus may over-represent successful proof-of-concept demonstrations and under-represent failure cases or practical limitations. Second, *reproducibility* is constrained by limited access to high-fidelity O&M simulators and proprietary data (e.g. SCADA/CM and maintenance logs) and incomplete reporting of experimental details (e.g. hyperparameters, reward shaping, baseline tuning, random seeds). These issues complicate rigorous replication and can make cross-paper comparisons sensitive to implementation choices rather than underlying algorithmic differences. Third, *generalizability* is limited because most reviewed studies evaluate DRL in stylized simulation environments with simplified wake, metocean, and logistics assumptions; consequently, reported gains may not transfer to real offshore operations or to farms with different layouts, failure modes, contractual constraints, and accessibility regimes. Finally, while many insights into DRL modelling choices (e.g. partial observability remedies, multi-agent coordination, multi-level actions) are transferable beyond offshore wind, their effectiveness may vary substantially across domains and should not be assumed without domain-specific validation.

Appendix A: Nomenclature and abbreviations

A2C	Advantage actor critic
A3C	Asynchronous advantage actor critic
CNN	Convolutional neural network
DCMAC	Deep centralized multi-agent actor critic
DDPG	Deep deterministic policy gradient
DDQN	Double deep Q network
DQN	Deep Q network
DRL	Deep reinforcement learning
GCN	Graph convolutional network
IOHMM	Input–output hidden Markov model
MDP	Markov decision process
O&M	Operations and maintenance
PHM	Prognostics and health management
POMDP	Partially observable Markov decision process
PPO	Proximal policy optimization
QMIX	Multi-agent value-based RL algorithm
RL	Reinforcement learning
RUL	Remaining useful life
SAC	Soft actor critic
SCADA	Supervisory control and data acquisition
W-QMIX	Weighted QMIX

Data availability. No underlying research data are associated with this study. This manuscript is a literature review and does not report original experimental, observational, or simulation data generated by the authors. Accordingly, no data repository or dataset archive applies to this work.

Author contributions. MB: conceptualization, data curation, formal analysis, investigation, methodology, software, validation, visualization, writing – original draft, writing – review and editing. XJ: supervision, writing – review and editing. RRN: supervision, writing – review and editing.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. The authors bear the ultimate responsibility for providing appropriate place names. Views expressed in the text are those of the authors and do not necessarily reflect the views of the publisher.

Financial support. This research has been supported by the Nederlandse Organisatie voor Wetenschappelijk Onderzoek (HoliDOCTOR project, grant no. KICH1.ED02.20.004).

Review statement. This paper was edited by Yolanda Vidal and reviewed by three anonymous referees.

References

Aafif, Y., Chelbi, A., Mifdal, L., Dellagi, S., and Majdouline, I.: Optimal preventive maintenance strategies for a wind turbine gearbox, Energy Reports, 8, 803–814, https://doi.org/10.1016/j.egyr.2022.07.084, 2022.

Abbas, A.: A Hierarchical Framework for Interpretable, Safe, and Specialised Deep Reinforcement Learning, Doctoral thesis, Technological University Dublin, https://doi.org/10.21427/p05p-az54, 2024.

Abbas, A. N., Chasparis, G. C., and Kelleher, J. D.: Hierarchical framework for interpretable and specialized deep reinforcement learning-based predictive maintenance, Data Knowl. Eng., 149, 102240, https://doi.org/10.1016/j.datak.2023.102240, 2024.

Abkar, M., Zehtabiyan-Rezaie, N., and Iosifidis, A.: Reinforcement learning for wind farm flow control: Current state and future actions, Renew. Energ., 205, 271–289, https://doi.org/10.1016/j.renene.2023.01.001, 2023.

Adadi, A. and Berrada, M.: Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI), IEEE Access, https://doi.org/10.1109/ACCESS.2018.2870052, 2018.

Andriotis, C. P. and Papakonstantinou, K. G.: Deep reinforcement learning driven inspection and maintenance planning under in-

complete information and constraints, Reliab. Eng. Syst. Safe., 212, 107551, https://doi.org/10.1016/j.ress.2021.107551, 2021.

Borsotti, M., Negenborn, R., and Jiang, X.: Model predictive control framework for optimizing offshore wind O&M, in: Advances in Maritime Technology and Engineering, CRC Press, 533–546, https://doi.org/10.1201/9781003508762-65, 2024.

Borsotti, M., Negenborn, R., and Jiang, X.: A review of multi-horizon decision-making for operation and maintenance of fixed-bottom offshore wind farms, Renew. Sust. Energ. Rev., 226, 116450, https://doi.org/10.1016/j.rser.2025.116450, 2026.

Bui, V. and Hollweg, G. V.: A Critical Review of Safe Reinforcement Learning Techniques in Smart Grid Applications, arXiv [preprint], https://doi.org/10.48550/arXiv.2409.16256, 2024.

Carroll, J., McDonald, A., Dinwoodie, I., McMillan, D., Revie, M., and Lazakis, I.: Availability, operation and maintenance costs of offshore wind turbines with different drive train configurations, Wind Energy, 20, 361–378, https://doi.org/10.1002/we.2011, 2017.

Chatterjee, J. and Dethlefs, N.: Scientometric review of artificial intelligence for operations & maintenance of wind turbines: The past, present and future, Renew. Sust. Energ. Rev., 144, 111051, https://doi.org/10.1016/j.rser.2021.111051, 2021.

Chen, M., Kang, Y., Li, K., Li, P., and Zhao, Y.-B.: Deep reinforcement learning for maintenance optimization of multi-component production systems considering quality and production plan, Qual. Eng., 1–12, https://doi.org/10.1080/08982112.2024.2373362, 2024.

Cheng, J., Liu, Y., Li, W., and Li, T.: Deep reinforcement learning for cost-optimal condition-based maintenance policy of offshore wind turbine components, Ocean Eng., 283, 115062, https://doi.org/10.1016/j.oceaneng.2023.115062, 2023.

Civera, M. and Surace, C.: Non-Destructive Techniques for the Condition and Structural Health Monitoring of Wind Turbines: A Literature Review of the Last 20 Years, Sensors-Basel, 22, 1627, https://doi.org/10.3390/s22041627, 2022.

Do, P., Nguyen, V.-T., Voisin, A., Iung, B., and Neto, W. A. F.: Multi-agent deep reinforcement learning-based maintenance optimization for multi-dependent component systems, Expert Syst. Appl., 245, 123144, https://doi.org/10.1016/j.eswa.2024.123144, 2024.

Dong, W., Zhao, T., and Wu, Y.: Deep Reinforcement Learning Based Preventive Maintenance for Wind Turbines, in: 2021 IEEE 5th Conference on Energy Internet and Energy System Integration (EI2), 2860–2865, https://doi.org/10.1109/EI252483.2021.9713457, 2021.

Dulac-Arnold, G., Levine, N., Mankowitz, D. J., Li, J., Paduraru, C., Gowal, S., and Hester, T.: Challenges of Real-World Reinforcement Learning: Definitions, Benchmarks and Analysis, Mach. Learn., 110, 2419–2468, https://doi.org/10.1007/s10994-021-05961-4, 2021.

Fox, H., Pillai, A. C., Friedrich, D., Collu, M., Dawood, T., and Johanning, L.: A Review of Predictive and Prescriptive Offshore Wind Farm Operation and Maintenance, Energies, 15, https://doi.org/10.3390/en15020504, 2022.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S.: Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor, arXiv [preprint], https://doi.org/10.48550/arxiv.1801.01290, 2018.

Hausknecht, M. and Stone, P.: Deep Recurrent Q-Learning for Partially Observable MDPs, in: AAAI Fall Symposium Series, arXiv, https://doi.org/10.48550/arxiv.1507.06527, 2015.

Huang, J., Chang, Q., and Arinez, J.: Deep reinforcement learning based preventive maintenance policy for serial production lines, Expert Syst. Appl., 160, 113701, https://doi.org/10.1016/j.eswa.2020.113701, 2020.

Igl, M., Zintgraf, L., Le, T. A., Wood, F., and Whiteson, S.: Deep Variational Reinforcement Learning for POMDPs, in: Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 80, PMLR, 2117–2126, http://proceedings.mlr.press/v80/igl18a.html (last access: 28 October 2025), 2018.

Jenkins, B., Prothero, A., Collu, M., Carroll, J., McMillan, D., and McDonald, A.: Limiting Wave Conditions for the Safe Maintenance of Floating Wind Turbines, J. Phys. Conf. Ser., 2018, https://doi.org/10.1088/1742-6596/2018/1/012023, 2021.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R.: Planning and acting in partially observable stochastic domains, Artif. Intell., 101, 99–134, https://doi.org/10.1016/S0004-3702(98)00023-X, 1998.

Kazemian, I., Yildirim, M., and Ramanan, P.: Attention is All You Need to Optimize Wind Farm Operations and Maintenance, arXiv, https://doi.org/10.48550/arxiv.2410.24052, 2024.

Kerkkamp, D., Bukhsh, Z., Zhang, Y., and Jansen, N.: Grouping of Maintenance Actions with Deep Reinforcement Learning and Graph Convolutional Networks, in: Proceeding of the 14th International Conference on Agents and Artificial Intelligence, vol. 2, SciTePress Digital Library, 574–585, https://doi.org/10.5220/0000155600003116, 2022.

Lee, J. and Mitici, M.: Deep reinforcement learning for predictive aircraft maintenance using probabilistic Remaining-Useful-Life prognostics, Reliab. Eng. Syst. Safe., 230, 108908, https://doi.org/10.1016/j.ress.2022.108908, 2023.

Lee, N., Woo, J., and Kim, S.: A deep reinforcement learning ensemble for maintenance scheduling in offshore wind farms, Appl. Energ., 377, https://doi.org/10.1016/j.apenergy.2024.124431, 2025.

Li, Q., Lin, T., Yu, Q., Du, H., Li, J., and Fu, X.: Review of Deep Reinforcement Learning and Its Application in Modern Renewable Power System Control, Energies, 16, https://doi.org/10.3390/en16104143, 2023.

Liang, J., Miao, H., Li, K., Tan, J., Wang, X., Luo, R., and Jiang, Y.: A Review of Multi-Agent Reinforcement Learning Algorithms, Electronics, 14, https://doi.org/10.3390/electronics14040820, 2025.

Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., and Wierstra, D.: Continuous Control with Deep Reinforcement Learning, in: International Conference on Learning Representations (ICLR), arXiv, https://doi.org/10.48550/arxiv.1509.02971, 2016.

Lowe, R., Wu, Y., Tamar, A., Harb, J., Pieter Abbeel, O., and Mordatch, I.: Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments, arXiv [preprint], https://doi.org/10.48550/arXiv.1706.02275, 2017.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D.:

Human-level control through deep reinforcement learning, Nature, 518, 529–533, https://doi.org/10.1038/nature14236, 2015.

Mnih, V., Badia, A., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K.: Asynchronous Methods for Deep Reinforcement Learning, Proceedings of Machine Learning Research, arXiv, https://doi.org/10.48550/arXiv.1602.01783, 2016.

Narayanan, S.: Reinforcement Learning in Wind Energy: A Review, Int. J. Green Energy, 20, 443–465, https://doi.org/10.1080/15435075.2023.2281329, 2023.

National Renewable Energy Laboratory: Offshore Wind Energy Market Assessment 2022, National Renewable Energy Laboratory, https://www.nrel.gov/wind/offshore-market-assessment.html (last access: 28 October 2025), 2022.

Nguyen, V.-T., Do, P., Voisin, A., and Iung, B.: Weighted-QMIX-based Optimization for Maintenance Decision-making of Multi-component Systems, in: Proceedings of the European Conference of the PHM Society 2022, vol. 7, 360–367, https://doi.org/10.36001/phme.2022.v7i1.3319, 2022.

Ogunfowora, O. and Najjaran, H.: Reinforcement and deep reinforcement learning-based solutions for machine maintenance planning, scheduling policies, and optimization, J. Manuf. Syst., 70, 244–263, https://doi.org/10.1016/j.jmsy.2023.07.014, 2023.

Pandit, R. and Wang, J.: A comprehensive review on enhancing wind turbine applications with advanced SCADA data analytics and practical insights, IET Renew. Power Gen., 18, 722–742, https://doi.org/10.1049/rpg2.12920, 2024.

Parisotto, E., Song, H. F., Rae, J. W., Pascanu, R., Gulcehre, C., Jayakumar, S. M., Jaderberg, M., Lopez Kaufman, R., Clark, A., Noury, S., Botvinick, M. M., Heess, N., and Hadsell, R.: Stabilizing Transformers for Reinforcement Learning, in: Proceedings of the 37th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 119, PMLR, arXiv [preprint], https://doi.org/10.48550/arXiv.1910.06764, 2020.

Pesántez, G., Guamán, W., Córdova, J., Torres, M., and Benalcazar, P.: Reinforcement Learning for Efficient Power Systems Planning: A Review of Operational and Expansion Strategies, Energies, 17, 2167, https://doi.org/10.3390/en17092167, 2024.

Pinciroli, L., Baraldi, P., Ballabio, G., Compare, M., and Zio, E.: Deep Reinforcement Learning Based on Proximal Policy Optimization for the Maintenance of a Wind Farm with Multiple Crews, Energies, 14, https://doi.org/10.3390/en14206743, 2021.

Qing, Y., Tong, Y., Qi, Z., and Li, Y.: A Survey on Explainable Reinforcement Learning: Concepts, Algorithms, and Challenges, arXiv [preprint], https://doi.org/10.48550/arXiv.2211.06665, 2022.

Rashid, T., Samvelyan, M., de Witt, C. S., Farquhar, G., Foerster, J. N., and Whiteson, S.: QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning, CoRR, arXiv, https://doi.org/10.48550/arXiv.1803.11485, 2018.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O.: Proximal Policy Optimization Algorithms, CoRR, arXiv, https://doi.org/10.48550/arxiv.1707.06347, 2017.

Stetco, A., Dinmohammadi, F., Zhao, X., Robu, V., Flynn, D., Barnes, M., Keane, J., and Nenadic, G.: Machine learning methods for wind turbine condition monitoring: A review, Renew. Energ., 133, 620–635, https://doi.org/10.1016/j.renene.2018.10.047, 2019.

Su, J., Huang, J., Adams, S., Chang, Q., and Beling, P. A.: Deep multi-agent reinforcement learning for multi-level preventive maintenance in manufacturing systems, Expert Syst. Appl., 192, 116323, https://doi.org/10.1016/j.eswa.2021.116323, 2022.

Sutton, R. S. and Barto, A. G.: Reinforcement Learning: An Introduction, 2nd edn., MIT Press, Cambridge, MA, ISBN 978-0-262-03924-6, http://incompleteideas.net/book/the-book-2nd.html (last access: 28 October 2025), 2018.

Tautz-Weinert, J. and Watson, S.: Using SCADA data for wind turbine condition monitoring – a review, IET Renew. Power Gen., 11, 382–394, https://doi.org/10.1049/iet-rpg.2016.0248, 2017.

Tusar, M. I. H. and Sarker, B. R.: Maintenance cost minimization models for offshore wind farms: A systematic and critical review, Int. J. Energ. Res., 46, 3739–3765, https://doi.org/10.1002/er.7425, 2022.

Valet, A., Altenmüller, T., Waschneck, B., May, M. C., Kuhnle, A., and Lanza, G.: Opportunistic maintenance scheduling with deep reinforcement learning, J. Manuf. Syst., 64, 518–534, https://doi.org/10.1016/j.jmsy.2022.07.016, 2022.

Vermeer, N.-J., Sørensen, J., and Crespo, A.: Wind turbine wake aerodynamics, Prog. Aerosp. Sci., 39, 467–510, https://doi.org/10.1016/S0376-0421(03)00078-2, 2003.

Wang, S., Vidal, Y., and Pozo, F.: Recent advances in wind turbine condition monitoring using SCADA data: A state-of-the-art review, Reliab. Eng. Syst. Safe., 267, 111838, https://doi.org/10.1016/j.ress.2025.111838, 2026.

Wei, S., Jin, X., Bao, Y., and Li, H.: Reinforcement Learning in Maintenance of Civil Infrastructures, in: Proceedings of the 36th International Conference on Machine Learning (ICML 2019), Workshop on Reinforcement Learning for Real Life (RL4RealLife), https://proceedings.mlr.press/v97/ (last access: 28 October 2025), 2019.

Williams, R. J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning, Mach. Learn., 8, 229–256, https://doi.org/10.1007/BF00992696, 1992.

Zhang, C., Li, Y.-F., and Coit, D. W.: Deep Reinforcement Learning for Dynamic Opportunistic Maintenance of Multi-Component Systems With Load Sharing, IEEE T. Reliab., 72, 863–877, https://doi.org/10.1109/TR.2022.3197322, 2023.

Zhang, E., Shen, F., Liu, S., Chen, G., Zhang, F., and Li, S.: Offshore wind power digital twin modeling system for intelligent operation and maintenance applications, E3S Web Conf., 546, https://doi.org/10.1051/e3sconf/202454602010, 2024.

Zhang, N. and Si, W.: Deep reinforcement learning for condition-based maintenance planning of multi-component systems under dependent competing risks, Reliab. Eng. Syst. Safe., 203, 107094, https://doi.org/10.1016/j.ress.2020.107094, 2020.

Zhao, F. J. and Zhou, Y.: Wind Farm Maintenance Scheduling Using Soft Actor-Critic Deep Reinforcement Learning, in: 2022 Global Reliability and Prognostics and Health Management (PHM-Yantai), 1–6, https://doi.org/10.1109/PHM-Yantai55411.2022.9942116, 2022.