Applied Energy 285 (2021) 116355

Contents lists available at ScienceDirect

Applied Energy

journal homepage: www.elsevier.com/locate/apenergy

Intelligent hurricane resilience enhancement of power distribution systems via deep reinforcement learning

Nariman L. Dehghani, Ashkan B. Jeddi, Abdollah Shafieezadeh *

Risk Assessment and Management of Structural and Infrastructure Systems (RAMSIS) Lab, Department of Civil, Environmental, and Geodetic Engineering, The Ohio State University, Columbus, OH, USA

HIGHLIGHTS

- The interplay of life-cycle resilience and time dependent reliability is introduced.
- A deep reinforcement learning framework for grid hardening decisions is proposed.
- A novel risk-based ranking is developed and integrated into reinforcement learning.
- A large-scale power distribution system consisting of over 7000 poles is studied.
- Resilience is improved by over 30% for a 100-year horizon.

ARTICLE INFO

Keywords:
Power distribution systems
Hardening strategies
Hurricanes
Life-cycle resilience
Deep Reinforcement Learning

ABSTRACT

Power distribution systems are continually challenged by extreme climatic events. The reliance of the energy sector on overhead infrastructures for electricity distribution has necessitated a paradigm shift in grid management toward resilience enhancement. Grid hardening strategies are among effective methods for improving resilience. Limited budget and resources, however, demand for optimal planning for hardening strategies. This paper develops a planning framework based on Deep Reinforcement Learning (DRL) to enhance the long-term resilience of distribution systems using hardening strategies. The resilience maximization problem is formulated as a Markov decision process and solved via integration of a novel ranking strategy, neural networks, and reinforcement learning. As opposed to targeting resilience against a single future hazard – a common approach in existing methods – the proposed framework quantifies life-cycle resilience considering the possibility of multiple stochastic events over a system's life. This development is facilitated by a temporal reliability model that captures the compounding effects of gradual deterioration and hazard effects for stochastic hurricane occurrences. The framework is applied to a large-scale power distribution system with over 7000 poles. Results are compared to an optimal strategy by a mixed-integer nonlinear programming model solved using Branch and Bound (BB), as well as the strength-based strategy by U.S. National Electric Safety Code (NESC). Results indicate that the proposed framework significantly enhances the long-term resilience of the system compared to the NESC strategy by over 30% for a 100-year planning horizon. Furthermore, the DRL-based approach yields optimal solutions for problems that are computationally intractable for the BB algorithm.

1. Introduction

Resilience enhancement of the power grid against natural hazards, especially extreme climatic events in coastal regions, is a critical step in modernizing the grid. Extreme weather-related hazards, such as hurricanes, have been responsible for over 80% of outages in the U.S. [1]. About 90% of these outages are attributed to failures in power distribution systems [2]. The increasing reliance of the society on electric energy as well as the high vulnerability of distribution systems to extreme climatic events have highlighted the need for resilience strategies that improve the ability of the grid to absorb shocks and rapidly recover from disruptive events. Operational and hardening measures are two main categories of resilience solutions [3]. Many studies investigated operational measures such as using microgrids (*e.g.,* [4–6]) and managing the grid recovery (*e.g.,* [7,8]). For example, Ding et al. [4]

* Corresponding author.
E-mail address: shafieezadeh.1@osu.edu (A. Shafieezadeh).

https://doi.org/10.1016/j.apenergy.2020.116355
Received 19 August 2020; Received in revised form 2 November 2020; Accepted 11 December 2020
Available online 12 January 2021

Nomenclature

Abbreviations

RL	reinforcement learning
DRL	deep reinforcement learning
MDP	Markov decision process
POMDP	partially observable Markov decision process
BB	branch and bound
DP	dynamic programming
LP	linear programming
NLP	nonlinear programming
MILP	mixed integer linear programming
MINLP	mixed integer nonlinear programming
DQN	deep Q-network
A2C	advantage actor-critic
A3C	asynchronous advantage actor-critic
DDPG	deep deterministic policy gradient
DNN	deep neural networks
ReLU	rectified linear unit
NESC	U.S. national electric safety code
RAW	risk achievement worth
EO	expected outage
EOR	expected outage reduction
MEOR	modified expected outage reduction
DGM	dichotomized Gaussian method

Variables and parameters

$a \in A$	action
$s \in S$	state
$r \in R$	reward
s_i, a_i, r_i	state, action, and reward at step$i$
T	total time steps in the planning horizon
γ	discount factor, penalty to uncertainty of future rewards,$0 < \gamma \leq 1$
J	return, discounted cumulative future rewards
$P(s^{'}, r\|s, a)$	transition probability
$\pi_\theta(a\|s)$	stochastic policy function, parametrized byθ
θ_π	parameters of policy function, weights of policy network
θ_V	parameters of value function, weights of value network
$Q(s,a)$	action-value function
$V(s)$	state-value function
$A(s,a)$	advantage value function
α	learning rate of gradient descent in policy network
β	learning rate of gradient descent in value network
D	experience replay memory of observations
N	number of samples of mini-batches of transition
M	number of episodes
C	target network update time step interval
T_{LC}	life-cycle of the system
t	time
t_i	time at step$i$
δt	time interval between two consecutive planning instances
$Q(t)$	performance of the system
$N_0(t)$	number of customers without power
N_t	total number of customers
R	expected resilience
t_c	control time
F_t	failure of a pole at time$t$
S_t	survival of a pole at time$t$
Γ_t	time to the most recent preventive replacement
H	height of pole
A_C	conductor area
v	wind speed
f_V	probability density function of wind speed
θ	wind direction
f_Θ	probability density function of wind direction
a_P	age of pole
f_A	probability density function of poles' age
N_C	number of clusters
LCR	life-cycle resilience
N_P	total number of utility poles in the system
$\mathbf{x}$	vector of decision variables
$\boldsymbol{I}$	identity matrix
TR	limit on the total constraint
PR	limit on the periodic constraint

investigated microgrids with various types of distributed generators and studied their effects on the resilience of power distribution systems. Wu and Sansavini [5] proposed a resilience-based microgrid design model that optimizes the placement and capacity of distributed energy resources and new power lines. A recent survey of studies that investigated microgrids as a potential solution for resilience enhancement of power distribution systems is provided in [6]. In a different approach to improving resilience, Arif et al. [7] proposed a two-stage stochastic Mixed-integer Linear Programming (MILP) model to optimize repair crew routing in distribution systems. In another study, Lei et al. [8] developed an optimization model for the logistics of the disaster response with the objective of shortening the recovery time. A few studies also explored hardening measures, such as replacement or reinforcement of utility poles (*e.g.*, [9–11]) and vegetation management (*e.g.*, [10,12]). For instance, Yuan et al. [9] developed an optimization model for hardening of power lines and placement of distributed generation units. Focusing on pole upgrades and vegetation management, Ma et al. [10] proposed an optimization model to identify plans that can effectively improve the resilience of power distribution systems. Moreover, Kuntz et al. [12] investigated the impact of trimming or removing trees that are close to overhead distribution lines and developed an optimal vegetation management plan that is aimed at improving the reliability of distribution systems. A recent review of different resilience enhancement strategies can be found in [13].

Among grid hardening solutions, replacement of poles is a key strategy that can mitigate disruptions by reducing the likelihood of pole failures relative to deteriorated poles [10,14]. Although hardening is an effective strategy, especially in regions that are prone to extreme weather events, it is costly [3]. Thus, budget and resource constraints have compelled the energy sector to prioritize pole replacements for hardening distribution lines. The U.S. National Electric Safety Code (NESC) [15] requires replacing poles once their remaining strength falls below two-thirds of the initial strength. This strategy is followed by over 90% of U.S. states as a hardening solution to increase the hazard reliability of power distribution systems [16]. Furthermore, risk-based prioritization strategies have been proposed as an alternative to NESC's strength-based strategy. For example, Salman et al. [17] presented a framework based on Risk Achievement Worth (RAW) to determine effective pole hardening prioritization strategies. These approaches, however, are not formulated as an optimization problem, therefore limiting the effectiveness of the strategies. Recently, the scope of optimal grid management has been extended from improving capacity or reducing risks to enhancing system resilience. Yuan et al. [9] formulated resilience enhancement of power systems against natural disasters using a two-stage robust optimization model. Ma et al. proposed a tri-level robust optimization [10] and a two-stage stochastic mixed-integer programming model [11] for enhancing the extreme weather resilience of power distribution systems. Lin and Bie [18]

presented a tri-level mixed integer optimization model to achieve resilience in the face of natural disasters. These studies, among others, are concerned with resilience to a single extreme event. An emerging paradigm in managing long-lived infrastructures is to maintain system resilience over extended horizons, as advocated by the U.S. National Infrastructure Advisory Council [19]. A major challenge toward this objective is that power distribution systems are not only exposed to extreme climatic stressors, but also undergo gradual aging deterioration. Concurrent impacts of extreme events and deterioration over long horizons where the system may be exposed to a stochastic series of extreme climatic events lead to a very challenging planning problem.

To analyze the long-term performance of the aging grid in the face of risks from natural hazards, the authors have introduced the notion of life-cycle resilience [20]. Life-cycle resilience refers to the capability of a system to absorb shocks and to quickly recover from incidents that may occur over the life-cycle of a system. This risk-based metric evaluates the long-term ability of the aging system under all possible hazard events considering uncertainties in the intensity and sequence of disruptive events. As the state of the grid infrastructure evolves with aging and corrective maintenance (*e.g.*, replacing damaged poles), preventive maintenance (*e.g.*, hardening) must be performed over time to maintain the capability of the system in facing hazards. Consequences of decisions at the time the action is taken, however, may not be immediately evident. Therefore, hardening decisions should be formulated as an optimal planning problem to ensure their long-term optimality. For large systems, arriving at optimal long-term decisions becomes a large-scale optimization problem with often complex objective functions. An optimal planning problem is commonly formulated as a combinatorial optimization problem, which is often solved by exact methods or heuristics. Although exact methods guarantee finding optimal solutions, they may not be applicable to large problems as the computational cost of these methods exponentially increases with size [21]. This limitation has increased interests in heuristics [22]. Heuristics trade off the optimality for computational time. These algorithms require substantial problem-specific research to guarantee proper performance [23,24]. Recently, methods based on Reinforcement Learning (RL) that tackle combinatorial optimization by systematically and automatically learning heuristics for combinatorial problems have gained attention. To solve a combinatorial optimization problem using RL algorithms, one must reformulate the problem as a sequential decision-making process [25]. Bello et al. [24] tackled solving the traveling salesman problem using Asynchronous Advantage Actor-Critic (A3C) method. Kool et al. [23] used the REINFORCE algorithm to solve several combinatorial optimization problems such as the traveling salesman, vehicle routing, and orienteering problems. Khalil et al. [26] used value-based RL algorithms to solve optimization problems over graphs. A recent survey on the application of RL for combinatorial optimization can be found in [25]. Recently, RL algorithms have been implemented in management and control problems in power systems including energy management [27,28], voltage control [29,30], and maximizing the profitability of aging grids [31]. More specifically, Du and Li [27] applied a model-free RL method to optimize retail electricity prices for distribution system operators by maximizing profitability, while minimizing peak-to-average ratio. Kong et al. [28] formulated the pricing problem of a service provider as a Markov Decision Process (MDP) and used the Q-learning algorithm to solve this problem. Yang et al. [29] proposed a framework based on Deep Q-Learning for minimizing the deviation of bus voltage from corresponding nominal values. Kou et al. [30] formulated the optimal voltage control in active distribution networks as a constrained MDP. They added a safety layer on top of the deep deterministic policy gradient (DDPG) algorithm to determine the optimal voltage control policy. Focusing on maintenance planning, Rocchtta et al. [31] studied the application of the Deep Q-Learning algorithm on a scaled-down aging power grid to maximize the profitability of the system.

This study presents a new framework based on Deep Reinforcement Learning (DRL) for life-cycle resilience enhancement of aging power distribution systems facing stochastic extreme events. Here, life-cycle resilience is enhanced by identifying the optimal long-term replacement plan for wood utility poles. In this context, resilience maximization is formulated as a sequential decision-making problem. Analysis of resilience over the life-cycle of power distribution systems requires tracking the reliability of utility poles over time. The interplay of life-cycle resilience of these systems and reliability of their components over time is introduced for the first time in this study. To track the reliability of utility poles, a recursive model is developed that incorporates uncertainties in the future state of poles arising from the compounding effects of aging deterioration and hurricane effects. Moreover, the resulting long-term non-stationary sequential decision-making problem is tackled using a novel risk-based ranking strategy integrated with the Advantage Actor-Critic (A2C) algorithm [32]. The proposed ranking strategy is built on risk-based metrics that can properly classify utility poles according to their contribution to the life-cycle resilience of the systems. The performance of the optimal hardening strategies derived by the proposed framework for a large power distribution system consisting of over 7,000 poles is compared with two other methods, including: (a) the strategy identified by a Mixed-Integer Nonlinear Programming (MINLP) model solved using the Branch and Bound (BB) algorithm and (b) the strength-based hardening strategy by NESC.

2. Methodology

2.1. General optimization model

To effectively enhance the resilience of power distribution systems over an extended horizon, we present a resilience-based optimization model where the objective is to maximize the expected life-cycle resilience of power systems via hardening strategies. The optimization model searches for the optimal planning policy *i.e.*, the optimal preventive maintenance actions for all poles in the distribution system over the entire planning horizon. Thus, maintenance decisions at each planning period are considered as decision variables. There are two possible actions per pole, including either replacing the pole with a new identical pole or doing nothing. Therefore, the decision variables are binary. Moreover, to incorporate budget and resource limits, two constraints are defined including periodic and total constraints. The former limits the number of pole replacements in each planning period, while the latter restricts the total number of pole replacements in the entire planning horizon. As these constraints are problem-specific, more details including the practical aspects are provided in Section 3. The general form of the proposed optimization model is as follows:

$$\begin{aligned} &\max_{\mathbf{x}} \ \mathbb{E}[LCR] \\ &s.t. \qquad \mathbf{x}_i \in \{0,1\}^{N_P}, \qquad i = 0, \cdots, T-1 \\ &\qquad \sum_{i=0}^{T-1} \boldsymbol{I}\mathbf{x}_i \le TR, \\ &\qquad \boldsymbol{I}\mathbf{x}_i \le PR, \qquad i = 0, \cdots, T-1 \end{aligned} \tag{1}$$

where $\boldsymbol{LCR}$ denotes the life-cycle resilience. $\boldsymbol{T}$ indicates the total periods (*i.e.,* time steps) in the planning horizon. $\mathbf{x}_i$ is the vector of decision variables at time step $\boldsymbol{i}$. $\boldsymbol{N_P}$ indicates the total number of utility poles in the system and $\boldsymbol{I}$ is the identity matrix of size $\boldsymbol{N_P}$. $\boldsymbol{TR}$and $\boldsymbol{PR}$ denote the limits for the total and period constraints, respectively. The goal is to find the optimal policy $\boldsymbol{\pi}^*$ that maximizes $\mathbb{E}[\boldsymbol{LCR}]$, where the policy $\boldsymbol{\pi} = [\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_{T-1}]$ is a matrix of size $\boldsymbol{N_P} \times \boldsymbol{T}$. The details of the life-cycle resilience calculation and the application of RL in solving this optimization model are provided in Section 2.2 and 2.3, respectively.

2.2. Life-cycle resilience quantification

A resilient power distribution system is able to absorb shocks from

extreme events, *e.g.*, hurricanes, and recover quickly. Uncertainties in the occurrence of hurricanes and their intensities, subsequent pole failures, and the recovery process necessitate a probabilistic treatment of resilience. To this end, a few studies proposed probabilistic resilience measures for power distribution systems (*e.g.* [33,34]). These measures focus on the short-term resilience of power for a specific disruptive event or a single uncertain hazard. However, critical infrastructure systems, such as power distribution networks, are intended to serve for very long periods. As a result, they may experience multiple extreme events during extended horizons. The objective of this study is to quantify and optimally enhance the long-term resilience of power distribution systems. Unlike conventional resilience assessments, long-term resilience, here referred to as life-cycle resilience, evaluates the performance of a system in the face of a multiplicity of stochastic disruptive events over extended horizons. Fig. 1 presents a schematic for the performance of a system facing the risk of extreme events over its lifetime. During the life-cycle of a system (*i.e.*, the period of $[0\ T_{LC}]$), the total number of extreme events, the time of hazard occurrences, the intensity of events, hazard-induced damages to the infrastructure, and the recovery processes are all uncertain. According to Fig. 1, for an extreme event, the time of occurrence is uncertain (point a). Point b indicates the uncertainty in the intensity of the extreme event and its immediate effects on the grid. The uncertainty around the recovery process after the extreme event is shown at point c. It is worth noting that the duration of the recovery time is often very short in reality compared to the entire life-cycle of a system. For each extreme event, the expected resilience is the area under the expected performance curve between the time of hazard occurrence and a control time (t_c) that is considered 30 days in this study.

A challenge in life-cycle resilience analysis is the quantification of the failure probability of overhead structures, which not only depends on design characteristics of the infrastructure and hazard intensities, but also on the state of the infrastructure prior to hazard occurrence. Whereas in short-term resilience assessments, the state of the system is fully known or known with high certainty, for long-term resilience quantification, this state is not known a priori. To elaborate this challenge in life-cycle resilience assessment and to explain the interplay of infrastructure resilience and reliability, a hypothetical but realistic scenario for the expected performance of a power distribution system and the expected reliability of utility poles in that system is illustrated in Fig. 2. In this scenario, it is assumed that two hazards occur at time t_{h1} and t_{h2}. Fig. 2(a) indicates that the system performs without any disruption until facing the first extreme event where the expected performance of the system drops at time t_{h1}. The system undergoes a recovery process until it reaches the fully functional condition. The system undergoes similar processes when it faces the second event at time t_{h2}. Fig. 2(b) shows the reliability – defined as the probability of survival – of two utility poles in the same system. In the considered scenario, both poles survive under the first extreme event. Pole A receives a preventive maintenance action at time t_m, whereas Pole B does not receive any preventive maintenance. The expected reliability of both poles decreases with time due to gradual aging deterioration and experiencing the first extreme event. At time t_m, Pole A is replaced with a new pole, thus, its reliability increases to nearly one since the new pole has not experienced any gradual deterioration or shock-type hazards. At time t_{h2} both poles in the system experience the second extreme event. In this scenario, Pole A resists against the hurricane shock due to its high reliability. Although high reliability does not guarantee survival of poles under a stochastic event, Pole A survives in this scenario as it has a high chance of survival. With similar analogy, Pole B fails under the second extreme event due to its low reliability. According to Fig. 2, after t_{out} from the second event, Pole B is replaced with a new pole with high reliability. This pole replacement illustrates a corrective maintenance action since the pole has already failed and should be replaced with a new one in order to recover the distribution system to its fully functional state. If another extreme event occurs after replacing Pole B, the new pole may resist against the stressor because its failure probability has significantly decreased after the replacement.

The above scenario indicates that life-cycle resilience quantification, gradual aging deterioration as well as preventive and corrective maintenance actions can influence the state of the overhead infrastructure prior to hazard occurrences. According to Fig. 2, the expected reliability of poles changes as time passes. It is crucial to keep track of the reliability of components over time to be able to estimate the expected life-cycle resilience. The impacts of preventive maintenance actions on the reliability of poles can be determined knowing the time of preventive replacement. The main challenge in projecting the reliability of components into the future arises from the impacts of gradual aging deterioration and corrective maintenance actions. These effects are discussed further in Section 2.2.2. Another important observation from Fig. 2 is that the reliability of components can decrease with time while the performance of the system remains at 100% when the system is not experiencing any extreme event. However, the decrease in reliability may increase the vulnerability of the system, therefore posing an increasing risk to the resilience of the system against future events.

The remainder of this section presents a framework for quantifying the life-cycle resilience of power distribution systems and an approach to estimate failure probability of utility poles within the framework. The process of probabilistic resilience quantification is illustrated in Fig. 3.

2.2.1. Probabilistic hurricane hazard model

As natural hazards, such as hurricanes, are region-specific, probabilistic hazard models should be developed specific to the region of interest. These models should probabilistically describe hazard

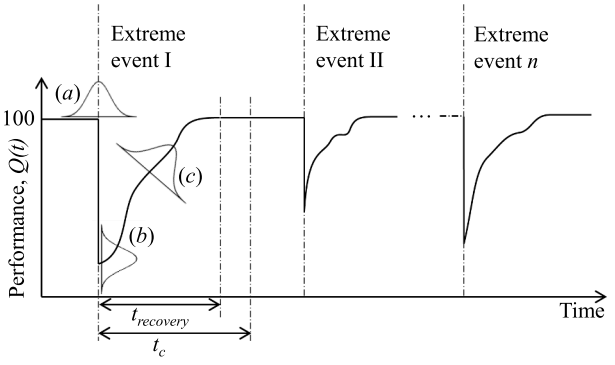

Fig. 1. A schematic illustration of the hazard performance of a system over its life-cycle.

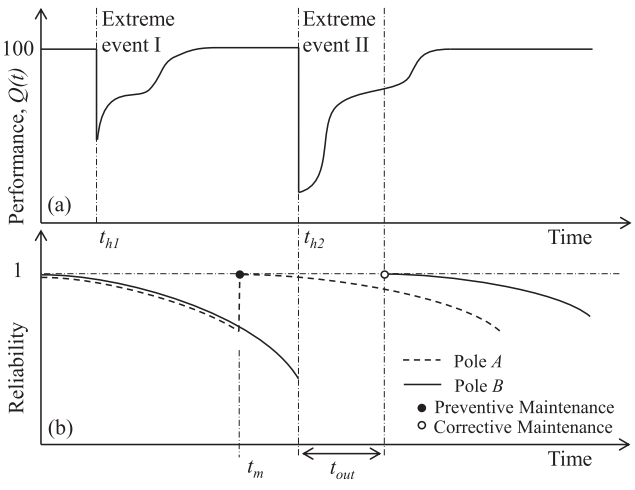

Fig. 2. The interplay of resilience and reliability: (a) the expected performance of a power distribution system over time and (b) the expected reliability of two hypothetical components in that system over time.

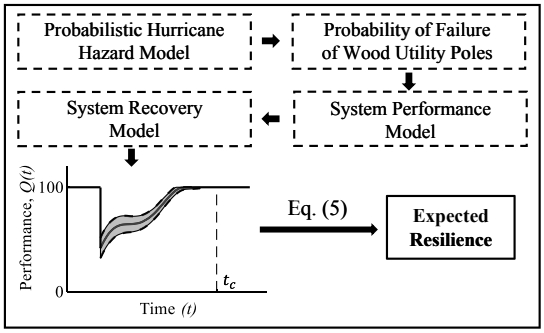

Fig. 3. Flowchart of the probabilistic resilience quantification.

characteristics that are significant for the infrastructure performance, *e.g.,* wind speed and wind direction. Probabilistic models for these hurricane characteristics are often derived based on historical data. For example, Peterka and Shahid [35] found that a Weibull distribution provides a reasonable fit to the hurricane wind speed in the southern U. S. Furthermore, Darestani [36] showed that the annual wind speed (in m/s) in Harris County, Texas, U.S. follows a Weibull distribution with scale parameter of 0.02 and shape parameter of 1.2. The studied distribution system in this investigation is assumed to be located in Harris County, thus, the probabilistic models provided in [36] are used here to probabilistically model the hurricane hazard. It is worth noting that here the hazard characteristics are considered to be stationary with time; however, these factors can be potentially affected by climate change, which is not considered in this study.

2.2.2. Probability of failure of wood utility poles

Hardening the grid via replacing poles is a preventive maintenance as it is implemented prior to pole failures. On the other hand, replacing damaged poles with new poles is regarded as corrective maintenance. For long planning horizons, it is anticipated that corrective maintenance actions will take place to recover the system from hazard-induced damages. These actions affect the age of poles, which plays a key role in the degradation state of overhead structures and therefore their failure probability. As the failure event of poles under a given hazard scenario is uncertain and the characteristics and the time of hazards are also uncertain, when the poles will fail over the planning horizon is not known a priori. Therefore, the age of poles in a distribution system is a stochastic process. To properly estimate failure probability of poles over time, an analytical model is introduced that integrates multi-dimensional fragility models with a recursive formulation.

(a) Multi-dimensional fragility model

Characteristics of wood utility poles, such as class, age, and height are significant factors for the failure probability of poles in the system. Moreover, the failure of poles under hurricanes depends on wind speed and wind direction. Failure probabilities of wood utility poles for a single hurricane conditional on the hazard and key characteristics of poles are determined using the multi-dimensional fragility models developed by Darestani and Shafieezadeh in [37]. These models provide failure probability of Class One through Class Seven Southern Yellow Pine wood poles conditional on wind speed and wind direction, given the conductor area as well as the class, age, and height of poles. Conductor area is calculated as the product of the conductor diameter, span length, and the number of conductors carried by poles. Details of these fragility models can be found in [37].

(b) Time-dependent failure probability of poles

The uncertainty in the age of poles over the planning horizon stems from the possibly needed corrective maintenance in the future. The impact of this uncertainty on the failure probability is quantified by considering the probability of all possible corrective maintenance actions. In this context, the failure probability of a pole at time t (*i.e.,* the time since the beginning of the planning horizon) is decomposed into two components. The first term corresponds to the probability that the pole fails at time t given that the pole has survived in all the previous times since the latest preventive maintenance. The second term represents the probability of failure of the pole at time t given that the pole has experienced at least one failure since the most recent preventive maintenance action. Following this decomposition, the probability of failure of each pole at time t is derived using the total probability theorem and probability chain rule as follows:

$$P(F_t|\Gamma_t,H,A_C,v,\theta)=P(F|a_P=\Gamma_t,H,A_C,v,\theta)\left[\prod_{l=1}^{min(t,\Gamma_t)}P\left(S_{t-l}|S_{t-l-1},\cdots,S_{t-min(t,\Gamma_t)},\Gamma_l,H,A_C\right)\right]+\sum_{k=1}^{min(t,\Gamma_t)}\left\{P(F|a_P=k,H,A_C,v,\theta)\left[\prod_{l=1}^{k-1}P(S_{t-l}|S_{t-l-1},\cdots,S_{t-k},\Gamma_l,H,A_C)\right]P(F_{t-k}|\Gamma_{t-k},H,A_C)\right\} \quad (2)$$

where Γ_t denotes the time to the most recent preventive replacement, H the height of the pole, A_C the conductor area, and v and θ the wind speed and wind direction, respectively. $P(F|a_P=\Gamma_t,H,A_C,v,\theta)$ can be obtained using the multi-dimensional fragility model in [37]. In Eq. (2), $\prod_{l=1}^{\min(t,\Gamma_t)}P\left(S_{t-l}|S_{t-l-1},\cdots,S_{t-\min(t,\Gamma_t)},\Gamma_l,H,A_C\right)$ indicates the probability that the pole has survived since the time of the most recent preventive maintenance action that was applied to this pole up to time $t-1$.Herein,Γ_l indicates the interval between time l and the time that the latest preventive maintenance is applied to the pole. It is worth noting that Γ_0 is equal to the age of the pole at the beginning of the planning horizon. However, Γ and age may not be the same during the rest of the planning horizon because age is impacted by both corrective and preventive maintenance actions, whereas Γ only represents the time to the latest preventive maintenance of the pole. In Eq. (2), $P(F_{t-k}|\Gamma_{t-k},H,A_C)$ denotes the probability of failure of the pole at time $t-k$, which can be determined as follows:

$$P(F_t|\Gamma_t,H,A_C)=\iint P(F_t|\Gamma_t,H,A_C,v,\theta)f_V(v).f_\Theta(\theta).dvd\theta \quad (3)$$

where f_V and f_Θ are the probability density functions of wind speed and wind direction, respectively. Using Eq. (3), the probability of survival of a pole at time t can be found as $P(S_t|\Gamma_t,H,A_C)=1-P(F_t|\Gamma_t,H,A_C)$.

2.2.3. System performance model

Performance of a power distribution system after a disruptive event can be defined in terms of the number of power outages. Herein, it is assumed that power outages occur only due to the failure of utility poles. A distribution system often operates as a radial system in which each node is connected to a single substation through a unique path. If a fault

or short circuit occurs in the system (*e.g.*, a utility pole fails such that the conductors break), protective devices cut off the power to downstream branches to prevent damage. Subsequently, a power distribution system reconfiguration is performed by opening and closing switches to provide power for some of the nodes that have lost power. The reconfiguration process in not considered in this study. Thus, the number of power outages is estimated as the number of nodes (*i.e.*, utility poles) that are not connected to any substation after a hurricane event.

When the probability of failure of all utility poles in the system is estimated, a set of random failure or survival scenarios for poles is generated to estimate the number of outages due to a hurricane event. The Dichotomized Gaussian Method (DGM) [38] is used in this study to generate correlated failure or survival scenarios because failure events of adjacent utility poles are correlated.

2.2.4. System recovery model

Following a disruptive event, faults in the distribution systems are identified and repair crews are dispatched to repair or replace damaged components and restore the system. In the aftermath of a hurricane, many utility poles may fail, thus, an effective recovery model is needed to prioritize repair and restoration in order to restore power in the entire system as fast as possible. In this study, it is assumed that the replacement of failed poles is prioritized based on the number of power outages that they have caused. The time to replace a failed pole is considered to follow a normal distribution with a mean of 5 h and a standard deviation of 2.5 h [34]. The repair time of a damaged conductor follows a normal distribution with a mean of 4 and a standard deviation of 2 h [34].

2.2.5. Probabilistic resilience quantification

For a specific hurricane event (*i.e.,* a specific wind speed and wind direction), the failure probability of all poles in the distribution system is estimated based on the procedure laid out in Section 2.2.2. The recovery model is subsequently applied and the state of recovery in terms of remaining power outages is tracked. The temporal performance of the power distribution system (Q) is estimated as:

$$Q(t) = 100\left(1 - \frac{N_o(t)}{N_t}\right) \tag{4}$$

where $N_o(t)$ is the number of customers without power at time t and N_t is the total number of customers in the power system. To account for the uncertainty in the performance of the system from the beginning of the hazard event to the end of the recovery time, this procedure is repeated for a large number of realizations of uncertain variables such as wind speed and wind direction. Subsequently, the expected resilience of the system is defined as the normalized area under the performance curve of the system as follow:

$$R = \frac{\int_0^{t_c} \mathbb{E}[Q(t)]dt}{t_c} \tag{5}$$

where $t_c$is the control time and is set as 30 days in this study. This control time is sufficiently long to ensure the complete restoration of the power in the entire studied distribution system.

2.3. Life-cycle resilience enhancement framework

The maximization problem in Section 2.1 is reformulated to conform with the RL framework and is subsequently solved via integration of a novel risk-based ranking and a DRL algorithm. In the remaining of this section, a brief overview of RL particularly the implemented A2C algorithm [32] is presented. Finally, the components of the RL methodology for the life-cycle resilience enhancement problem are formalized.

2.3.1. Reinforcement learning

MDP provides a mathematical framework for optimal sequential decision-making in problems with finite sets of discrete-state and discrete-action spaces. In MDP settings, Dynamic Programming (DP), Linear Programming (LP), and Nonlinear Programming (NLP) are employed to determine optimal policies. These methods, however, face significant computational challenge for large-scale, multi-state, and multi-component systems [39]. RL is an alternative to DP, LP, and NLP methods for long-term sequential decision-making, especially with recent advancements in deep neural networks that have enabled the application of RL to high-dimensional problems. In RL, the decision-maker, known as the agent, interacts with the system, called the environment, at state $s_i \in S$ by taking actions $a_i \in A$ at each time step i. The agent learns a mapping from states to actions based on the new information she receives from the environment ($s_{i+1} \sim P(s_i, a_i)$) and a scalar reward value ($r_i \sim R(s_i, a_i)$). These environments are formalized through MDPs with the 4-tuple (S, A, P, R). The goal of the agent is to maximize the cumulative reward by exploring the environment and following the policies learned based on the gathered information and eventually to identify the optimal policy in this process. Several classes of RL techniques have been developed and applied, as briefly reviewed next.

2.3.1.1. Q-Learning. In the classical Q-learning algorithm [40], the agent constructs an action-value function, $Q(s, a)$ that represents the quality of each action in every possible state. Then, the agent updates the Q-function by interacting with the environment through taking actions and receiving rewards in order to achieve the optimal policy ($\pi^*(s_i)$). However, for having the full representation of action-value function, the agent should explore every possible state-action combination. In the case of large systems with a large number of states and actions, the process of calculating action-value function can be computationally demanding.

To alleviate the high computational cost of Q-learning methods, Deep Q-Network (DQN) has been introduced, which combines deep neural networks with Q-learning. DQN algorithms provide an estimate of the Q-function value for those areas of state space that are not explored by the agent. Besides the non-linear function approximation capabilities of DQNs, target network and experience replay are two key features that can further improve the stability and convergence guarantees of DQNs [41]. In particular, target Q-networks ($Q^{'}$) have a structure identical to the Q-network, but at time step interval C, the weights of the target network are updated by replacing the weights of the Q-function. On the other hand, experience replay introduces a memory of observations (D) and the Q-function is updated over a mini-batch sample of size N from memory D.

2.3.1.2. Policy gradient. Policy gradient methods directly search for the optimal policy in the policy space [42], offering advantage over Q-learning methods in cases where the action space is large or the Q-function is complex to learn. The objective function in policy gradient methods is the expected cumulative future reward:

$$J(\theta) = \mathbb{E}\left[\sum_{i=0}^{T-1} r_{i+1} \middle| \pi_\theta\right] \tag{6}$$

where r_{i+1} is the reward received by taking action a_i in state s_i, π_θ is the policy with parameters θ, and T is the total time steps in the planning horizon. Policy governs a_i in a certain state s_i, *i.e.*, $\pi_\theta(s_i) = a_i$. With this definition, the gradient of $J(\theta)$ with respect to the policy parameters θ can be calculated as:

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{i=0}^{T-1} \nabla_\theta log\pi_\theta(a_i|s_i) Q(s_i, a_i)\right] \tag{7}$$

where $Q(s_i, a_i)$ is the action-value or Q-function which is defined as $\mathbb{E}\left[\sum_{n=0}^{T-i-1} \gamma^n r_{i+n+1} | s_i, a_i\right]$, where $\gamma \in [0, 1]$ is the discount factor.

2.3.1.3. Advantage Actor-Critic. The third class of RL algorithms is

Actor-Critic methods which take advantage of both Q-learning and policy gradient techniques [43]. In the Actor-Critic method, the Critic function often estimates values of the state-value function (*i.e.,* V-function) via a neural network with weights θ_V and the Actor function updates the policy distribution via a neural network with weights θ_π [15]. Despite their high capabilities, the performance of Actor-Critic methods can degrade due to noisy gradients and high variance, among other factors [44]. These factors can lead to instability and slow convergence of Actor-Critic methods [44]. A2C [32] has been developed to reduce the variance and increase the stability of Actor-Critic methods by adding the state-value function as a baseline in the policy gradient.

In A2C, an advantage function (*i.e.,* $A(s_i, a_i)$) substitutes the Q-function in Eq. (7). This function is defined as the subtraction of the state-value function from the action-value function (*i.e.,* $Q(s_i, a_i) - V(s_i)$). Thus, the modified gradient of $J(\theta)$ can be derived as:

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\sum_{i=0}^{\mathcal{T}-1} \nabla_\theta log\pi_\theta(a_i|s_i)(Q(s_i, a_i) - V(s_i))\right] \tag{8}$$

Considering Bellman equations [45], the advantage value can be rewritten as:

$$A(s_i, a_i) = r_{i+1} + \gamma V(s_{i+1}|\theta_V) - V(s_i|\theta_V) \tag{9}$$

This modification to the objective function improves the stability of Actor-Critic methods by reducing the variance of the gradient. This study uses the A2C algorithm to identify optimal long-term non-stationary sequential decisions. Having the policy gradient values, weights of the policy and value networks can be updated recursively via gradient descent algorithm as:

$$\theta_\pi \leftarrow \theta_\pi + \alpha A(s_i, a_i)\nabla_{\theta_\pi} log\pi(s_i|\theta_\pi) \tag{10}$$

$$\theta_V \leftarrow \theta_V + \beta A(s_i, a_i)\nabla_{\theta_V} V(s_i|\theta_V) \tag{11}$$

where α and β are the learning rates in the gradient descent algorithm for policy and value networks, respectively. The workflow of the A2C algorithm is presented in Fig. 4.

2.3.2. Reinforcement learning encoding

To approach the life-cycle resilience maximization problem for a distribution system based on the RL framework, the environment, set of actions, reward function, and agent must be properly defined. As there are thousands of utility poles in a power distribution system, solving the optimization problem is computationally interactable. To this end, a novel risk-based ranking is proposed to overcome the curse of dimensionality by classifying the poles in a distribution system. Therefore, before discussing the RL components in the case of a power distribution system, the porposed risk-based method for ranking the utility poles is presented.

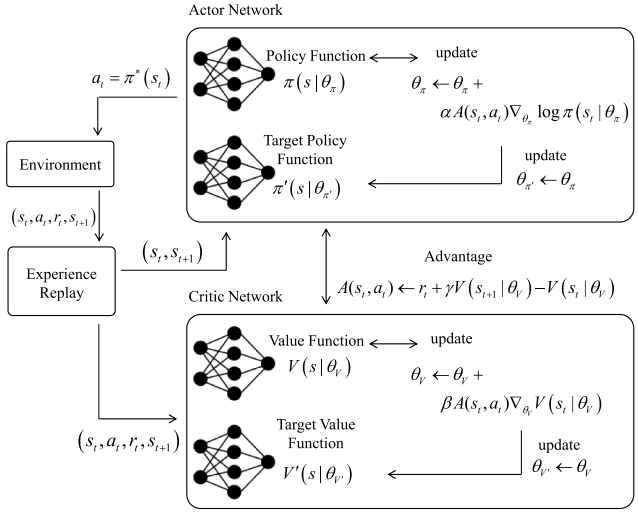

Fig. 4. Workflow of the A2C methodology.

2.3.2.1. Risk-based ranking.

To overcome the mentioned curse of dimensionality, the poles are ranked and classified into N_C clusters of equal sizes. Two risk-based indices are introduced for ranking the poles. The mathematical formulation, advantages and disadvantages of these indices are elaborated as follows:

(a) Expected outage reduction

Expected Outage Reduction (EOR) is a risk index that estimates the reduction in the expected power outages of a distribution system when an existing pole is replaced with a new one [36]. This index is calculated as follows:

$$EOR_{j,t} = N_j\left[P_j(F_t|\Gamma_t, H, A_C) - P_j(F_t|0, H, A_C)\right] \tag{12}$$

where $EOR_{j,t}$ indicates the direct expected reduction in the outages of the system if pole j is replaced with a new pole of age zero at time t.$P_j(F_t|\Gamma_t, H, A_C)$ is the probability of failure of pole j with the time to the most recent preventive replacement Γ_t, height H, and conductor area A_C. This probability can be computed using Eq. (3).$P_j(F_t|0, H, A_C)$ indicates the failure probability of pole j that is just replaced at time t. According to Eq. (12), EOR of each pole depends on time t. As the time step changes, the EOR of the poles may change. Thus, if the poles are ranked and classified based on EOR, their ranks may vary for different time steps. Consequently, as the time step changes, poles in a cluster may not belong to that cluster anymore. The time dependency of EOR limits the effectiveness of this risk index in ranking the poles for the purpose of classification. In fact, EOR can be effective for classification if the age of all poles in a system is the same at the beginning of the planning horizon.

(b) Modified expected outage reduction

To overcome the limitation of EOR in classifying poles, a novel risk-based metric is proposed here. This index called Modified Expected Outage Reduction (MEOR) can be determined as follows:

$$MEOR_j = N_j \iiint [P(F|a_P, H, A_C, v, \theta) - P(F|0, H, A_C, v, \theta)]f_V(v).f_\Theta(\theta).f_A(a_P).dv.d\theta.da_P \tag{13}$$

where $MEOR_j$ is the modified expected outage reduction of pole j, and $f_A(a_P)$ denotes the distribution of the age of poles. $P(F|a_P, H, A_C, v, \theta)$ is a conditional failure probability of a pole and is computed using the multi-dimensional fragility models provided in [37]. The main advantage of MEOR over EOR is the time independency of this risk index. Although MEOR is a proper risk index for ranking and classifying poles, it cannot be as effective as EOR for distribution systems with poles that have the same age.

2.3.2.2. Environment.

The environment in the RL algorithm is a power distribution system that is susceptible to gradual deterioration and stochastic occurrences of hurricane events. Herein, the environment is fully observable meaning that the environment state is the same as the state that is observed by the agent. Incorporating inspection in this planning problem requires extending the problem to a Partially Observable Markov Decision Process (POMDP) formulation. Moreover, it is assumed that the topology of the distribution system does not change over time. In small-size power distribution systems, the environment state at each time step, $s_i \in S$, can be set as the time to the most recent preventive replacement of utility poles. However, for real-size distribution systems which consist of many utility poles, the size of action and state spaces

becomes significantly large. To overcome the curse of dimensionality, in this framework, poles are ranked based on the proposed risk-based ranking index, and then classified into N_C clusters of equal sizes. Thus, the environment state at each time step, s_i, is a vector of size N_C and includes the mean of the time to the latest preventive replacement of poles per cluster. Possible actions in the environment as well as the reward at each time step are described next.

2.3.2.3. Actions. At each time step, i, replacement actions for hardening distribution poles can be performed on clusters. There are two possible actions for each cluster, including either replacing all poles in that cluster with new poles of the same class and height or doing nothing. Therefore, all poles in each cluster receive the same action. In this RL methodology, action is represented as a vector of size N_C with binary values corresponding to whether to replace each cluster or not.

2.3.2.4. Rewards. In RL, the agent learns by directly interacting with the environment and receiving rewards based on the impact of each action on the environment. Thus, the reward function is a key component in achieving a well-trained RL agent. A well-defined reward function helps the agent to converge faster to an optimal policy. As the objective is to enhance the resilience, the reward at each time step, r_i, should be related to the resilience of the system. However, an agent cannot enhance the resilience by replacing an unlimited number of utility poles. In fact, the utility companies are able to replace only a limited number of poles per year. As mentioned in Section 2.1, to account for these practical aspects, periodic and total constraints are defined in the planning for resilience enhancement of power distribution systems. If the agent violates these constraints, the achieved reward will be penalized with a large penalty. Two reward functions that are related to the expected resilience are introduced as follows:

(a) Reward as a function of expected outages

Expected Outages (EO) of a pole is the risk associated with the failure of that pole. The EO of poles is related to the resilience of the system since decreasing the EO of all poles can increase system resilience. This risk-based metric is calculated as follows:

$$EO_{j,i} = N_j \times P_j(F_{t_i}|\Gamma_{t_i}, H, A_C) \tag{14}$$

where $EO_{j,i}$ denotes the EO of pole j at time step i. $N_j$is the number of power outages that will occur in the entire distribution system due to the failure of pole j. $P_j(F_{t_i}|\Gamma_{t_i}, H, A_C)$ is the probability of failure of pole j with the time to the most recent preventive replacement Γ_{t_i}, height H, and conductor area A_C. This probability can be calculated using Eq. (3). It is worth noting that t_i is the time at time step i, and is determined by the product of i and δt, where δt is a constant value that indicates the time interval between two consecutive time steps (*e.g.*, $\delta t = t_{i+1} - t_i$). Reducing the total EO of the system can enhance system resilience. Thus, the reward at time step i is defined as:

$$r_i = \sum_{i=1}^{N_P} -log_{10}(EO_{j,i}) \tag{15}$$

(b) where N_P is the total number of poles in the system.

(c) Reward as a function of expected resilience

The second reward model is a direct function of resilience, which is computed as follows:

$$r_i = -log_{10}\left(1 - \frac{R_{t_i}}{100}\right) \tag{16}$$

where R_{t_i} denotes the expected resilience of the distribution system at time t_i. This formulation is used rather than the resilience itself in order to increase the sensitivity of expected resilience to effects of preventive actions.

2.3.2.5. Agent. At each time step i, the agent interacts with the environment by taking actions $a_i \in A$. Then, it observes state $s_i \in S$ and receives a scalar reward r_i. According to the MDP formulation, the state transition depends on the current state, s_i, and the taken action at that time step, a_i. In this framework, if the agent decides to replace a cluster of poles at time step i, the state of that cluster in the next time step, $i+1$, will be set to δt (*i.e.,* the time interval between two consecutive time steps). However, if the agent does not replace the cluster at time step i, the state of that cluster in the next time step will be increased by δt. The main goal of the agent is to increase the received rewards after executing actions. The continued interaction between the agent as the decision maker and the environment (*i.e.*, the power distribution system) leads to gradual improvement of the preventive replacement policy via the acquired experience.

The proposed resilience enhancement framework for power distribution systems is outlined in Algorithm 1. This framework focuses on life-cycle resilience enhancement of power distribution systems via effective pole replacement. Other hardening strategies such as vegetation management as well as operational measures including distributed generation can be investigated in the future for their effects on long-term resilience of the systems. Moreover, this framework is developed based on the assumption that the environment is fully observable. Therefore, the optimization problem is formulated as an MDP. However, the proposed framework can be extended to partially observable environments, where the problem should be reformulated as a POMDP.

Algorithm 1. (*Proposed Resilience Enhancement Framework.*)

1: Determine total time steps, T, in the planning horizon and the time interval between two consecutive time steps, δt
2: Determine the number of clusters, N_C, and perform risk-based ranking on utility poles
3: Initialize critic and actor networks with random weights θ_π & θ_V
4: Initialize target critic and actor networks with weights $\theta_{\pi'} = \theta_\pi$ & $\theta_{V'} = \theta_V$
5: Initialize hyperparameters, α in Eq. (10), β in Eq. (11), & min-batch size N
6: **for** episode $= 1 : M$
7: **for** time step $i = 1 : T$
8: Agent takes action a_i, which is a binary vector of size N_C
9: Agent observes s_{i+1}, which is a vector of size N_C
10: Agent receives scalar r_i using Eq. (15) or Eq. (16)
11: Store experience (s_i, a_i, r_i, s_{i+1}) in D
12: Sample mini-batch of size N from D
13: Update the critic network using Eq. (10)

(*continued on next page*)

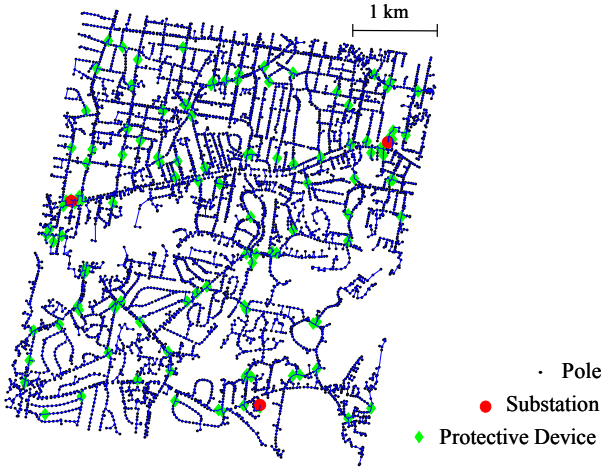

Fig. 5. The studied power distribution system topology, including 7051 wood utility poles, 115 protective devices, and three substations (courtesy of Darestani [36]).

(*continued*)

14: Update the actor network using Eq. (11)
15: Every C steps, update target networks $\theta_{\pi'} = \theta_{\pi}$ & $\theta_{V'} = \theta_{V}$
16: **end for**
17: **end for**
18: Return the optimal policy $\pi^*(s_i)$ as the hardening strategy

3. Results

The proposed resilience enhancement framework is applied to a power distribution system that is assumed to be in Harris County, Texas, US. The system consists of 7051 wood utility poles, 115 protective devices and three substations (Fig. 5). According to [36], the probability density function of annual wind speed (in m/s) in this region follows a Weibull distribution with scale parameter of 0.02 and shape parameter of 1.2. Optimal pole replacement strategies as hardening solutions for this system are determined for two cases:

I) Planning horizon of 60 years: Height, span length, and class of poles are selected based on a real-world case study, while the age of all poles is assumed to be 25 years at the beginning of the planning.

II) An extended horizon of 100 years: All features of the poles, including their age at the start of planning, are considered to be different representing real-world cases.

Although the studied distribution system is located in Harris County, Texas, the proposed framework is general and applicable to power distribution systems located in any geographic region.

In practice, long-term plans are divided into short-term planning periods (also known as operation plans) [46]. Thus, in Case 1 and 2, the long-term planning horizon is discretized into 20 and 33 short-term periods of three years (*i.e.,* δt=3), respectively. In other words, the replacement decisions are made every three years because short-term planning periods for distribution systems are often three years [46]. To tackle the curse of dimensionality, the 7051 poles are categorized into 15 clusters (*i.e.,* N_C=15) with almost equal sizes (~6.7% of the poles in the system per cluster). Due to the limited budget and resource, the number of pole replacements cannot exceed a threshold per period of decision-making. It is estimated that at least three percent of wood utility poles in the U.S. are replaced every year [47]. Thus, around 10% of the poles in a system may need to be replaced in every three-years period. However, the studied distribution system is in Harris County which is classified as a humid subtropical zone with high wood decay rate. Therefore, more than 3% of the poles in the system may need to be replaced every year. Accordingly, the maximum number of replacements per period (*i.e.,* three years) is set to ~13.3% of the poles in the system (*i.e.*, maximum two clusters per period).

3.1. Case I

As explained in Section 2.3.2, to overcome the curse of dimensionality, the poles are classified into clusters. Classifying the poles here is facilitated by the EOR risk-based index at the beginning of the planning horizon. In this example, the poles are classified into 15 clusters based on their initial EOR ranking, where clusters 1 and 15 have the lowest and highest initial EOR, respectively. As mentioned in Section 2.3.2.1, it may not be suitable to cluster the poles based on their initial EOR, but rather based on their current EOR. If a cluster is replaced at time step i, the age of all poles in the cluster becomes zero at that time step, thus, the EOR of poles in that cluster changes. Consequently, the ranking of poles may change and poles in a cluster may not belong to that cluster anymore. However, in this example, the age of all poles is the same at the beginning of planning, therefore, this limitation does not affect the classification of the poles. Following the risk-based ranking and formation of clusters, Steps 3–18 in Algorithm 1 are implemented to identify the optimal policy. EO in Eq. (15) is used here as the reward function. As the budget is limited, the number of pole replacements is constrained for each planning period as well as for the entire planning horizon. The former and latter constraints are called periodic and total constraints, respectively. As mentioned earlier, the periodic constraint is set to ~13.3% of the poles in the system (*i.e.*, maximum two clusters per period). The total constraint for the 60 years of the planning horizon is set at 7051 poles, which is equal to the total number of poles in the system. The outcome of the proposed framework is referred to as the DRL-based strategy. For implementing the DRL algorithm, we employ two deep neural networks (DNNs) for actor and critic, respectively (see Fig. 4). These DNNs have three Rectified Linear Unit (ReLU) hidden layers each with 128 hidden units. The Adam optimizer is used for learning the network weights. A discount factor of 0.99, a learning rate of 0.0007 for both actor and critic, and entropy regularization with a weight of 0.01 are used. The models are trained on a random mini-batch sample of size 20 from experience replay. The architecture and other hyperparameters of these neural networks are selected through controlled experiments.

The performance of the DRL-based strategy is compared to the strength-based pole replacement strategy of NESC (NESC strategy) as well as a pole replacement strategy obtained from a MINLP model (MINLP-based strategy) introduced by the authors in [20]. To make a fair comparison between these three hardening strategies, the total number of replacements is set at 7051 poles, equal to the total number of poles in the system. In addition, the total number of pole replacements per period is limited to ~13.3% of the poles in the system. According to the NESC strategy, the residual strength of the poles at each time step should be identified and those with over 33% loss of strength should be replaced in that time step. As shown by Shafieezadeh et al. [48], the residual strength of wood poles follows a lognormal distribution with age-dependent parameters. For the MINLP model, the reward function in Section 2.3.2.4.b is used as the objective function, and the periodic and total limits on pole replacements are considered as the constraints. This model is solved using LINDO [49] which uses convex relaxations and reformulations within a Branch and Bound (BB) algorithm to solve non-convex problems. To improve the computational time of solving the MINLP model and the proposed framework in finding optimal strategies, instead of direct calculation of Eq. (15), a surrogate model is used. In this surrogate model, the state of the environment is the input and the reward in Eq. (15) is the output. Solving a MINLP problem using the BB algorithm with LINDO requires an objective function with a mathematical expression. To this end, symbolic regression [50] is used here to develop the surrogate model. More details on the mathematical formulation of the MINLP model can be found in [20].

The DRL and MINLP-based strategies are presented in Fig. 6. The DRL-based planning recommends replacing the cluster with the highest initial EOR (cluster 15) twice, whereas no replacement is recommended for the cluster with the lowest EOR (cluster 1), indicating the effectiveness of the proposed risk-based ranking. On the other hand, the MINLP strategy recommends all clusters to be replaced once.

Fig. 7(a) and (b) present the cumulative number of pole replacements by each strategy and the expected resilience of the power distribution system over its life-cycle, respectively. According to Fig. 7 (a), the total number of replacements is the same for all strategies because the same constraint for the total number of replacements is applied to all cases. It is evident that the MINLP-based and DRL-based strategies significantly outperform the NESC strategy for almost all duration of the planning horizon for the same total number of pole replacements. According to Fig. 7(b), following the NESC strategy, the system will have the minimum expected resilience of 97.576% at year 45, while in the same year, the MINLP and DRL strategies yield an expected resilience of 99.768% and 99.773%, respectively. Although the difference of 2.19% between the resilience of the NESC strategy and the other two strategies may seem marginal, such improvements in the expected annual resilience of power distribution systems can save millions of dollars. For instance, Ouyang and Duenas-Osorio [34] showed that for

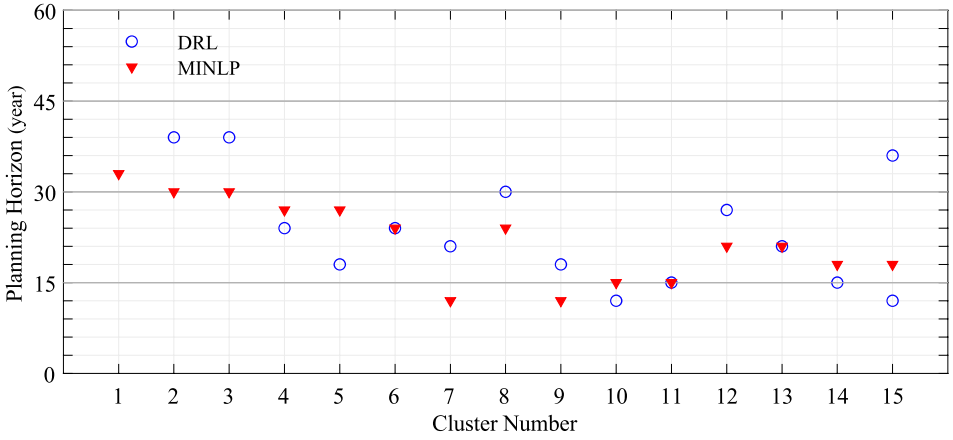

Fig. 6. Optimal hardening strategy as the solution of DRL and MINLP.

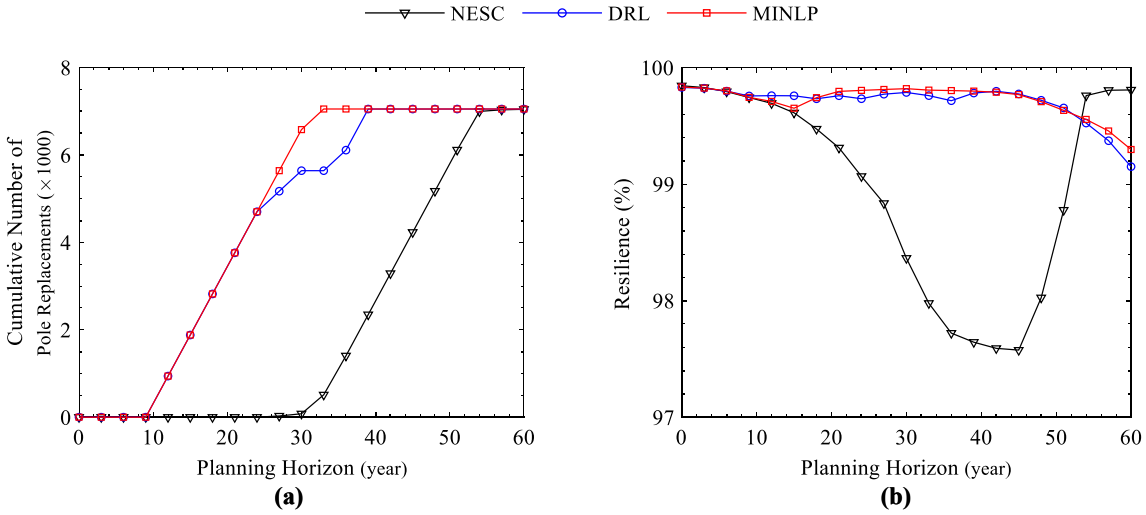

Fig. 7. Impacts of hardening strategies on (a) the cumulative number of pole replacements (b) the expected resilience of the power distribution system.

an electric power system in Harris County, Texas, a 0.038% decrease in the expected annual resilience can lead to an economic loss of up to 83 million dollars per year.

3.2. Case II

In this example, the proposed framework is applied to a real-world power distribution system in order to enhance its resilience over an extended horizon of 100 years. As explained in Section 2.3.2.1 and Section 3.1, classifying the poles based on EOR is effective when only for cases where the age of all utility poles is identical at the beginning of the planning period. In this example, the MEOR index is used to rank the poles. Calculating the MEOR for a pole requires the probability density function of the poles' age. According to [51], the age of poles in a distribution system similar to the one considered in this study follows a lognormal distribution with the mean and standard deviation of 31.2 years and 14.6 years, respectively. The poles are ranked based on MEOR and then classified into 15 clusters, with clusters 1 and 15 having the lowest and highest MEOR, respectively. Following risk-based ranking of the poles and forming the clusters, Steps 3–18 in Algorithm 1 are implemented to identify the optimal hardening strategy. In this example both EO in Eq. (15) and the expected resilience in Eq. (16) are used as the reward function. The corresponding optimal strategies are referred to as DRL(1) and DRL(2), respectively. It is worth reminding that in the process of training, the agent is penalized for actions that violate the periodic and total constraints. All the hyperparameters for implementing the A2C algorithm are the same as Case I, except that we employ DNNs for an actor and critic that have three hidden layers each with 256 hidden units. Considering the size of this problem relative to the problem in Case I, and based on the Universal Approximation Theorem, a neural network with a larger number of hidden units is chosen through a systematic experimentation [52].

In this example, the performance of DRL(1) and DRL(2) strategies is compared to the pole replacement strategy set by NESC. To make a fair comparison between these three hardening strategies, the total number of replacements is set to 14,102 poles for the entire planning horizon. This limit is selected because following the NESC approach, 14,102 poles should be replaced over 100 years. As mentioned earlier, in addition to this total constraint, a limit is also set on the number of pole replacements (*i.e.*, ~13.3% of the poles in the system) per period for the NESC and both DRL strategies. It should be noted that the planning horizon in Case II is longer than that in Case I, which further adds to the complexity of the problem. More specifically, the presented resilience enhancement problem in this study is inherently combinatorial in nature and it belongs to the set of NP-complete problems, meaning that the solution time increases exponentially as the problem size increases [53]. Thus, in this example, the results of the proposed framework are not compared to the solution of the MINLP model because the length of the planning horizon is nearly twice the previous example, and despite the efficiency of the BB algorithm in solving MINLP models, it fails to

converge to an optimal or sub-optimal solution for this problem. This observation is also compatible with other investigations on sequential decision-making problems using the BB algorithm (*e.g.,* [54]). This limitation is attributed to the formation of rooted trees in the BB algorithm for solving combinatorial optimization problems. At every step, the BB algorithm selects a branching rule on the node of the tree. This process significantly impacts the overall size of the tree, and consequently the computational demand of the algorithm. On the other hand, an RL agent with a proper learning setting can emulate the node selection policy proportional to the running time of the BB algorithm [25].

Although the desirable performance of RL algorithms in solving combinatorial optimization problems has been shown in recent studies (*e.g.,* [23,24,55]), their computational demand can be high if the calculation of the reward function is costly. To improve the computational time of the proposed DRL framework, instead of direct calculation of the reward function at each time step, two DNNs are developed to estimate the reward models in Eq. (15) and Eq. (16). The architecture of these DNN models consists of six hidden layers with 128, 64, 32, 16, 8, and 4 hidden units, respectively. The ReLU activation is adopted for all hidden layers. The Adam optimizer is applied to minimize the loss function. The DRL(1) and DRL(2) pole replacement policies are presented in Fig. 8. Both strategies recommend replacing clusters 14 and 15 three times, while replacing clusters 1 and 2 once in the entire planning horizon. This observation shows the capability of MEOR in properly classifying the poles, as clusters 1 and 2 include poles with the lowest MEOR, *i.e.* poles with the least importance, whereas clusters 14 and 15 consist of poles with the highest MEOR. According to Fig. 8, the interval between two consecutive replacements is about 40 to 50 years for poles in clusters 3 to 13. For clusters 14 and 15, the optimal time for replacing poles falls within the range of 30 to 35 years. This observation indicates that no replacement is needed for poles with an age less than 30 years. However, it should be noted that these conclusions depend on the periodic and total constraints set for pole replacements.

Fig. 9(a) and (b) show the cumulative number of pole replacements by each strategy and the expected resilience of the distribution system over its considered life-cycle, respectively. As seen in Fig. 9(a), the total number of replacements over the entire planning horizon is the same for the NESC and DRL strategies because the same constraint for the total number of replacements is applied to all cases. According to Fig. 9(b), DRL(1) and DRL(2) outperform the NESC strategy over the duration of interest for the same number of total pole replacements. As explained in Case I, although the enhancement in the expected resilience seems marginal, the achieved improvement can reduce the hurricane-induced economic losses by millions of dollars [34]. In the proposed resilience enhancement framework, the objective is to improve the resilience over the entire horizon meaning that the area under the curves in Fig. 9 should be maximized. Although DRL(1) and DRL(2) recommend different replacement policies (see Fig. 8), following these strategies leads to approximately the same area under the life-cycle resilience curve. Thus, both strategies are near-optimal solutions. This observation indicates that both proposed reward functions are appropriate. Therefore, considering that evaluating Eq. (15) is computationally less demanding compared to Eq. (16), we recommend using Eq. (15) as the reward function.

The differences between the expected resilience of the DRL strategies and the NESC strategy over the planning horizon are shown in Fig. 10. The maximum improvement by the DRL strategies is about 0.6%, which can significantly reduce the incurred costs of loss of power to communities in hazard prone regions [34]. Moreover, the DRL(1) and DRL(2) strategies improve the cumulative expected resilience of the power distribution system by 29.8% and 30.2% with respect to the NESC strategy over the 100 years period of the planning horizon. This result highlights the capability of the DRL-based approaches in offering optimal solutions that can significantly enhance the resilience of power systems.

4. Conclusion

Power distribution systems are among the most critical infrastructures as they supply electric energy for many societal and economic activities. Given the ever-increasing reliance of the society on electricity and the devastating short and long-term consequences of long-duration power outages caused by extreme climatic events, resilience enhancement of power distribution systems has become a key goal for the energy sector. This paper presented an approach based on Deep Reinforcement Learning (DRL) to optimally improve the long-term resilience of large-scale aging power distribution systems that are exposed to hurricane risks. The proposed framework is general and applicable to power distribution systems with different topologies and in any geographic region that is susceptible to hurricane hazards. It is shown that an agent empowered with proper learning models can optimally plan for the long service life of a system that faces the risk of multiple stochastic extreme events over an extended horizon. We proposed solving the resulting long-term sequential decision-making problem with the Advantage Actor-Critic (A2C) algorithm. The framework is applied to a large-scale power distribution system with over 7000 poles and the optimal preventive maintenance strategies are identified. The results were compared to an optimal strategy determined by a Mixed-Integer Nonlinear Programming (MINLP) model solved using a Branch and Bound (BB) algorithm as well as the strategy set by U. S. National Electric Safety Code (NESC). Both the proposed DRL and MINLP-based strategies were able to significantly enhance the resilience of the system compared to the NESC strategy. It is further shown that the DRL-based approach can yield optimal solutions for cases that are

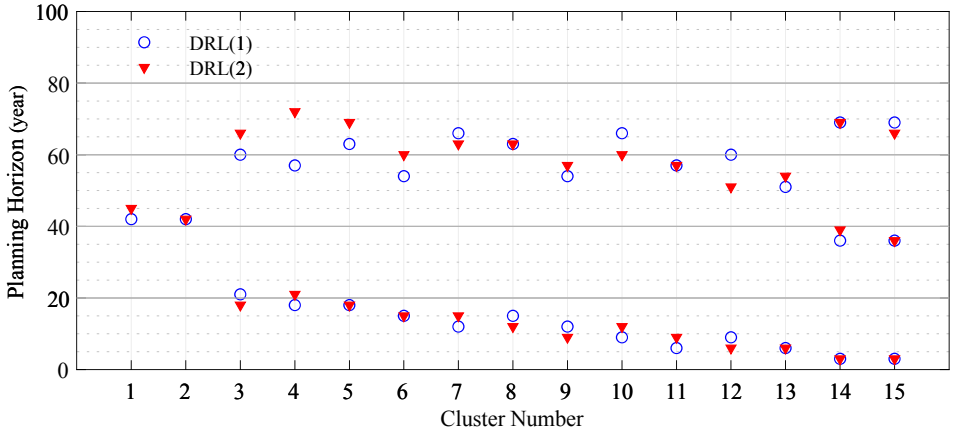

Fig. 8. Optimal hardening strategy as the solution of DRL(1) and DRL (2).

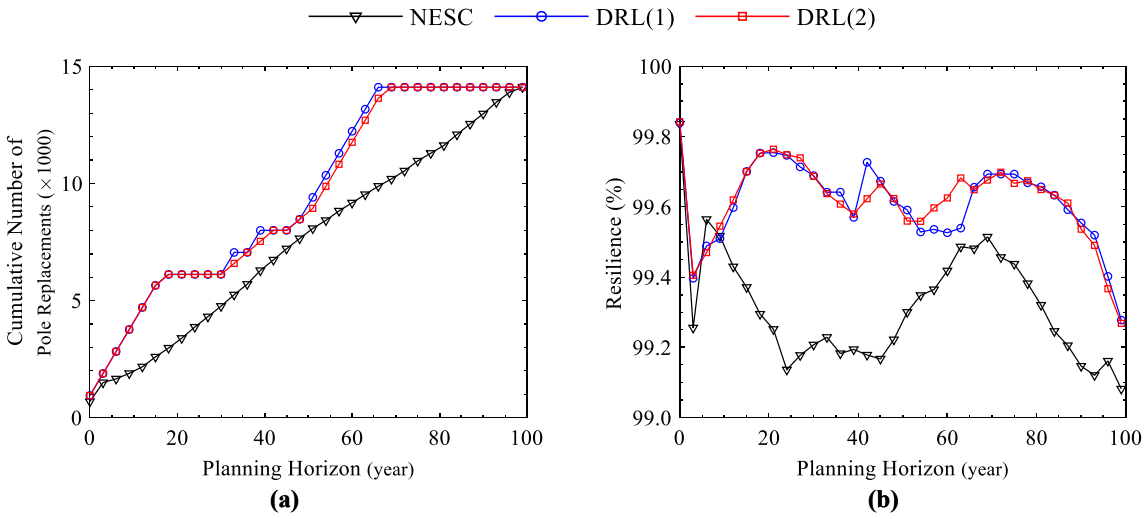

Fig. 9. Impacts of hardening strategies on (a) the cumulative number of pole replacements (b) the expected resilience of the power distribution system.

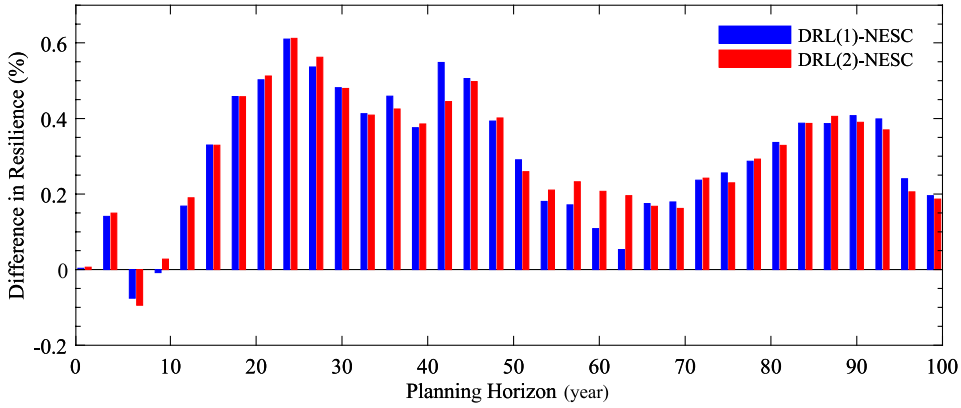

Fig. 10. Difference between the expected resilience of DRL strategies and the NESC strategy.

intractable for the BB algorithm. The proposed framework can be extended to include other resilience enhancement strategies, such as vegetation management and operational measures. Moreover, the proposed approach is based on the assumption that the environment is fully observable. This assumption can be relaxed in future research by extending the framework to partially observable environments.

CRediT authorship contribution statement

Nariman L. Dehghani: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft. **Ashkan B. Jeddi:** Conceptualization, Methodology, Software, Formal analysis, Writing - review & editing. **Abdollah Shafieezadeh:** Conceptualization, Methodology, Writing - review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was partially supported by the National Science Foundation under grants CMMI-1635569 and 2000156, as well as Lichtenstein endowment at The Ohio State University.

References

[1] Panteli M, Mancarella P. Influence of extreme weather and climate change on the resilience of power systems: Impacts and possible mitigation strategies. Electr. Power Syst. Res. 2015;127:259–70.
[2] Executive Office of the President, "Economic Benefits of Increasing Electric Grid Resilience to Weather Outages, IEEE USA Books & eBooks, 2013. [Online]. Available: https://www.energy.gov/sites/prod/files/2013/08/f2/Grid%20Resiliency%20Report_FINAL.pdf.
[3] Panteli M, Trakas DN, Mancarella P, Hatziargyriou ND. Power systems resilience assessment: Hardening and smart operational enhancement strategies. Proc. IEEE 2017;105(7):1202–13.
[4] Ding T, Lin Y, Bie Z, Chen C. A resilient microgrid formation strategy for load restoration considering master-slave distributed generators and topology reconfiguration. Appl. Energy 2017;199:205–16.
[5] Wu R, Sansavini G. Integrating reliability and resilience to support the transition from passive distribution grids to islanding microgrids. Appl. Energy 2020;272:115254.
[6] Hussain A, Bui V-H, Kim H-M. Microgrids as a resilience resource and strategies used by microgrids for enhancing resilience. Appl. Energy 2019;240:56–72.
[7] Arif A, Ma S, Wang Z, Wang J, Ryan SM, Chen C. Optimizing service restoration in distribution systems with uncertain repair time and demand. IEEE Trans. Power Syst. 2018;33(6):6828–38.
[8] Lei S, Chen C, Li Y, Hou Y. Resilient disaster recovery logistics of distribution systems: Co-optimize service restoration with repair crew and mobile power source dispatch. IEEE Trans. Smart Grid 2019;10(6):6187–202.
[9] Yuan W, Wang J, Qiu F, Chen C, Kang C, Zeng B. Robust optimization-based resilient distribution network planning against natural disasters. IEEE Trans. Smart Grid 2016;7(6):2817–26.
[10] Ma S, Chen B, Wang Z. Resilience enhancement strategy for distribution systems under extreme weather events. IEEE Trans. Smart Grid 2016;9(2):1442–51.
[11] Ma S, Su L, Wang Z, Qiu F, Guo G. Resilience enhancement of distribution grids against extreme weather events. IEEE Trans. Power Syst. 2018;33(5):4842–53.

[12] Kuntz PA, Christie RD, Venkata SS. Optimal vegetation maintenance scheduling of overhead electric power distribution systems. IEEE Trans. Power Deliv. 2002;17(4):1164–9.
[13] Jufri FH, Widiputra V, Jung J. State-of-the-art review on power grid resilience to extreme weather events: Definitions, frameworks, quantitative assessment methodologies, and enhancement strategies. Appl. Energy 2019;239:1049–65.
[14] Onyewuchi UP, Shafieezadeh A, Begovic MM, DesRoches R. A probabilistic framework for prioritizing wood pole inspections given pole geospatial data. IEEE Trans. Smart Grid 2015;6(2):973–9.
[15] National Electrical Safety Code® (NESC®), "ANSI/IEEE Standard," 2017.
[16] Trevor N. Bowmer, "National Electrical Safety Code (NESC) Update," presented at the ATIS Protection Engineers Group Conference, Dallas, Texas, USA, 1207, [Online]. Available: https://peg.atis.org/wp-content/uploads/sites/16/2018/08/NESC_Update_TrevorBowmer.pdf.
[17] Salman AM, Li Y, Stewart MG. Evaluating system reliability and targeted hardening strategies of power distribution systems subjected to hurricanes. Reliab. Eng. Syst. Saf. 2015;144:319–33.
[18] Lin Y, Bie Z. Tri-level optimal hardening plan for a resilient distribution system considering reconfiguration and DG islanding. Appl. Energy 2018;210:1266–79.
[19] Panteli M, Mancarella P. The grid: Stronger, bigger, smarter?: Presenting a conceptual framework of power system resilience. IEEE Power Energy Mag. 2015;13(3):58–66.
[20] Dehghani NL, Darestani YM, Shafieezadeh A. Optimal Life-Cycle Resilience Enhancement of Aging Power Distribution Systems: A MINLP-Based Preventive Maintenance Planning. IEEE Access 2020;8:22324–34.
[21] Ausiello G, Crescenzi P, Gambosi G, Kann V, Marchetti-Spaccamela A, Protasi M. Complexity and approximation: Combinatorial optimization problems and their approximability properties. Springer Science & Business Media; 2012.
[22] Lee KS, Geem ZW. A new meta-heuristic algorithm for continuous engineering optimization: harmony search theory and practice. Comput. Methods Appl. Mech. Eng. 2005;194(36–38):3902–33.
[23] W. Kool, H. Van Hoof, and M. Welling, "Attention, learn to solve routing problems!," ArXiv Prepr. ArXiv180308475, 2018.
[24] I. Bello, H. Pham, Q. V. Le, M. Norouzi, and S. Bengio, "Neural combinatorial optimization with reinforcement learning," ArXiv Prepr. ArXiv161109940, 2016.
[25] N. Mazyavkina, S. Sviridov, S. Ivanov, and E. Burnaev, "Reinforcement learning for combinatorial optimization: A survey," ArXiv Prepr. ArXiv200303600, 2020.
[26] Khalil E, Dai H, Zhang Y, Dilkina B, Song L. Learning combinatorial optimization algorithms over graphs. Adv. Neural Info. Process. Syst. 2017:6348–58.
[27] Du Y, Li F. Intelligent multi-microgrid energy management based on deep neural network and model-free reinforcement learning. IEEE Trans. Smart Grid 2019;11(2):1066–76.
[28] Kong X, Kong D, Yao J, Bai L, Xiao J. Online pricing of demand response based on long short-term memory and reinforcement learning. Appl. Energy 2020;271:114945.
[29] Yang Q, Wang G, Sadeghi A, Giannakis GB, Sun J. Two-timescale voltage control in distribution grids using deep reinforcement learning. IEEE Trans. Smart Grid 2019;11(3):2313–23.
[30] Kou P, Liang D, Wang C, Wu Z, Gao L. Safe deep reinforcement learning-based constrained optimal control scheme for active distribution networks. Appl. Energy 2020;264:114772.
[31] Rocchetta R, Bellani L, Compare M, Zio E, Patelli E. A reinforcement learning framework for optimal operation and maintenance of power grids. Appl. Energy 2019;241:291–301.
[32] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Harley, T. P. Lillicrap et al., "Asynchronous methods for deep reinforcement learning," in International conference on machine learning, 2016, pp. 1928–1937.
[33] Hosseini MM, Parvania M. Quantifying impacts of automation on resilience of distribution systems. IET Smart Grid 2020;3(2):144–52.
[34] Ouyang M, Duenas-Osorio L. Multi-dimensional hurricane resilience assessment of electric power systems. Struct. Saf. 2014;48:15–24.
[35] Peterka JA, Shahid S. Design gust wind speeds in the United States. J. Struct. Eng. 1998;124(2):207–14.
[36] Darestani YM. Hurricane Resilience Quantification and Enhancement of Overhead Power Electric Systems. Columbus, OH, USA: The Ohio State University; 2019.
[37] Darestani YM, Shafieezadeh A. Multi-dimensional wind fragility functions for wood utility poles. Eng. Struct. 2019;183:937–48.
[38] Emrich LJ, Piedmonte MR. A method for generating high-dimensional multivariate binary variates. Am. Stat. 1991;45(4):302–4.
[39] Andriotis CP, Papakonstantinou KG. Managing engineering systems with large state and action spaces through deep reinforcement learning. Reliab. Eng. Syst. Saf. 2019;191:106483.
[40] C. J. C. H. Watkins, "Learning from delayed rewards," 1989.
[41] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare et al., "Human-level control through deep reinforcement learning," nature, vol. 518, no. 7540, pp. 529–533, 2015.
[42] R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in Advances in neural information processing systems, 2000, pp. 1057–1063.
[43] T. Degris, M. White, and R. S. Sutton, "Off-policy actor-critic," ArXiv Prepr. ArXiv12054839, 2012.
[44] Grondman I, Busoniu L, Lopes GA, Babuska R. A survey of actor-critic reinforcement learning: Standard and natural policy gradients. IEEE Trans. Syst. Man Cybern. Part C Appl. Rev. 2012;42(6):1291–307.
[45] Bellman R. On the theory of dynamic programming. Proc. Natl. Acad. Sci. U.S.A. 1952;38(8):716.
[46] Kiessling F, Nefzger P, Nolasco JF, Kaintzyk U. Overhead power lines: planning, design, construction. Springer; 2014.
[47] Feldman J, Shistar TA, Poles Poison. A Report About Their Toxic Trail and the Safer Alternatives. National Coalition Against Misuse Pesticides 1997.
[48] Shafieezadeh A, Onyewuchi UP, Begovic MM, DesRoches R. Age-dependent fragility models of utility wood poles in power distribution networks against extreme wind hazards. IEEE Trans. Power Deliv. 2013;29(1):131–9.
[49] Lin Y, Schrage L. The global solver in the LINDO API. Optim. Methods Softw. 2009;24(4–5):657–68.
[50] Koza JR. Genetic programming: on the programming of computers by means of natural selection, vol. 1. MIT press; 1992.
[51] Darestani YM, Shafieezadeh A, DesRoches R. Effects of adjacent spans and correlated failure events on system-level hurricane reliability of power distribution lines. IEEE Trans. Power Deliv. 2017;33(5):2305–14.
[52] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, Deep learning, vol. 1. MIT press Cambridge, 2016.
[53] Floudas CA, Lin X. Mixed integer linear programming in process scheduling: Modeling, algorithms, and applications. Ann. Oper. Res. 2005;139(1):131–62.
[54] Moghaddam KS, Usher JS. Preventive maintenance and replacement scheduling for repairable and maintainable systems using dynamic programming. Comput. Ind. Eng. 2011;60(4):654–65.
[55] Yousefi N, Tsianikas S, Coit DW. Reinforcement learning for dynamic condition-based maintenance of a system with individually repairable components. Qual. Eng. 2020:1–21.