Optimization of the Operation and Maintenance of Renewable Energy Systems by Deep Reinforcement Learning

Luca Pinciroli[1], Piero Baraldi[1*], Guido Ballabio[2], Michele Compare[2], Enrico Zio[1,3]

[1]Energy Department, Politecnico di Milano, Milan, Italy.
[2]Aramis S.r.l., Milan, Italy.
[3]MINES ParisTech, PSL Research University, CRC, Sophia Antipolis, France.
[*]piero.baraldi@polimi.it – Energy Department, Politecnico di Milano, Via Lambruschini 4, 20156, Milan, Italy.

Optimization of the Operation and Maintenance of Renewable Energy Systems by Deep Reinforcement Learning

Luca Pinciroli[1], Piero Baraldi[1*], Guido Ballabio[2], Michele Compare[2], Enrico Zio[1,3]

[1]Energy Department, Politecnico di Milano, Milan, Italy.
[2]Aramis S.r.l., Milan, Italy.
[3]MINES ParisTech, PSL Research University, CRC, Sophia Antipolis, France.
[*]piero.baraldi@polimi.it – Energy Department, Politecnico di Milano, Via Lambruschini 4, 20156, Milan, Italy.

Abstract

Equipment of renewable energy systems are being supported by Prognostics & Health Management (PHM) capabilities to estimate their current health state and predict their Remaining Useful Life (RUL). The PHM health state estimates and RUL predictions can be used for the optimization of the systems Operation and Maintenance (O&M). This is an ambitious and challenging task, which requires to consider many factors, including the availability of maintenance crews, the variability of energy demand and production, the influence of the operating conditions on equipment performance and degradation and the long time horizons of renewable energy systems usage. In this work, we develop a novel formulation of the O&M optimization of renewable energy systems equipped with PHM capabilities as a sequential decision problem and we resort to Deep Reinforcement Learning (DRL) to solve it. The proposed solution approach combines Proximal Policy Optimization (PPO), as DRL algorithm, Imitation Learning (IL), for pre-training the learning agent, and a model of the environment which describes the renewable energy system behavior. The solution approach is tested by its application to a wind farm O&M problem. The optimal solution found is shown to outperform those provided by other DRL algorithms. Also, the approach does not require to select a-priori a maintenance strategy, such as corrective, scheduled, condition-based or predictive but, rather, it discovers the best performing policy by itself.

Keywords: Renewable Energy Systems, Wind farm, Operation and Maintenance, Prognostics and Health Management, Optimization, Deep Reinforcement Learning.

Acronyms and symbols

ACER	Actor-Critic with Experience Replay
AHP	Analytic Hierarchy Process
AI	Artificial Intelligence
CM	Corrective Maintenance
CPS	Cyber-Physical System
DNN	Deep Neural Network
DQN	Deep Q-Network
DRL	Deep Reinforcement Learning
ELECTRE	Elimination Et Choice Translating Reality
GA	Genetic Algorithm
IL	Imitation Learning
IoT	Internet of Things
IT	Information Technology
LCC	Life Cycle Cost
MCDM	Multiple Criteria Decision Making
O&M	Operation & Maintenance
PdM	Predictive Maintenance
PHM	Prognostics and Health Management
PM	Preventive Maintenance
PPO	Proximal Policy Optimization
RUL	Remaining Useful Life
RL	Reinforcement Learning
SDP	Sequential Decision Process
TOPSIS	Technique for Order Preference by Similarity to Ideal Solution
TPE	Tree-structured Parzen Estimator
TRPO	Trust Region Policy Optimization

WT	Wind Turbine
T_M	Time horizon
t	Generic decision time
N_{T_M}	Number of times a decision is taken
L	Number of units
l	Generic unit
Λ	Set of units
T_l	Ground-truth failure time of unit l
T_l^*	Failure time of unit l in nominal conditions
$f_{T_l^*}$	Probability density function of T_l^*
λ_f	Unit failure rate
ν_τ	Degradation factor in non-nominal conditions
R_l^*	RUL of unit l in nominal conditions
$\hat{R}_l^*$	Estimate of the RUL of unit l in nominal conditions
$\hat{\boldsymbol{R}}_t^*$	Vector containing the RUL estimates of the L units at time t
ϵ_R	Error of the RUL estimate
σ_R	Standard deviation of ϵ_R
Ag_l	Age of unit l
$\boldsymbol{Ag}_t$	Vector containing the ages of the L units at time t
P_l	Ground-truth power production of unit l
$\hat{P}_l$	Estimate of the power production
$\hat{\boldsymbol{P}}_t$	Vector containing the power production estimates of the L units at time t
ϵ_P	Error of the power production estimate
σ_P	Standard deviation of ϵ_P
J	Number of days for which the prediction of P is available
j	Generic prediction day
G_t	Revenues at time t
K	Maximum revenue per unit
H	Maintenance crew depot
Π_{CM}	Corrective maintenance downtime
Π_{PM}	Preventive maintenance downtime
μ_{CM}	Corrective maintenance repair rate
μ_{PM}	Preventive maintenance repair rate
U_{CM}	Corrective maintenance cost
U_{PM}	Preventive maintenance cost
MT_l	Time needed to complete the maintenance intervention of unit l
$\boldsymbol{MT}_t$	Vector containing the times to complete the maintenance interventions of the L units at time t
X_t	Costs at time t
$\mathcal{S}$	State space
s	Generic state
$\boldsymbol{s}_t$	State vector at time t
$\mathcal{A}$	Action Space
$\boldsymbol{A}$	Vector of possible actions
a	Generic action
a_t	Scalar representing the action of the maintenance crew at time t
$\mathcal{P}$	Transition probability
$\mathcal{R}$	Reward function
r_t	Reward at time t
γ	Discount factor
π	Generic policy
π^*	Optimal policy
$V^\pi(s)$	Value function
$Q^\pi(s,a)$	Action-value function
$\hat{A}^\pi(s,a)$	Advantage function estimate
F	Objective function
ϵ	PPO clipping hyperparameter
v_{cut-in}	Cut-in wind speed
v_{rated}	Rated wind speed
$v_{cut-out}$	Cut-out wind speed
L'	Number of units with inaccurate RUL predictions
l'	Generic unit with inaccurate RUL predictions
Λ'	Set of units with inaccurate RUL predictions
q_{PM}	Restoration factor

1. Introduction

In the last years, the interest of the energy industry on renewable sources of energy has grown significantly due to social, economic and environmental perspectives (Sanz-Bobi, 2014). A renewable energy plant requires, like any other energy production plant, an Operation and Maintenance (O&M) strategy, for ensuring the proper functioning of the plant's components, reducing the risk of failure, and increasing the production availability of the overall system.

The recent developments of Information Technology (IT) have enabled the possibility of equipment monitoring and direct communications between machines within a Cyber-Physical System (CPS) (Ustundag and Cevikcan, 2018). The implementation of this paradigm in the production and operation environments is often termed as Industry 4.0 (Tjahjono *et al.*, 2017), and exploits the combination of big data, Internet of Things (IoT), Cyber-Physical Systems and Artificial Intelligence (AI) to obtain environments where smart machines communicate with one another to enable the automation of production lines and the, monitoring, detection, elaboration of data and information for preventing equipment failures (Barreto, Amaral and Pereira, 2017). The final goal is not just to improve production management but also to effectively manage equipment and reduce downtime (Terrissa *et al.*, 2016).

In this context, Prognostics and Health Management (PHM) plays a leading role, using condition monitoring data for estimating the equipment health state and predicting its Remaining Useful Life (RUL), i.e., the remaining amount of time that a component can be operated before it loses its functional capabilities (Okoh *et al.*, 2014). Several algorithms for RUL prediction have been developed (Simões, Gomes and Yasin, 2011) and several successful applications to industrial components have been reported in literature (Kwon *et al.*, 2016; Al-Dulaimi *et al.*, 2019; Cai *et al.*, 2020). In particular, Predictive Maintenance (PdM) exploits PHM outcomes to set efficient, maintenance interventions, which aim at providing the right part to the right place at the right time, giving, therefore, the opportunity of maximizing system availability and minimizing the Life Cycle Cost (LCC) of the system and the losses (Compare, Baraldi and Zio, 2020).

Although the advantages of PdM are intuitive, the application of PdM to renewable energy systems should consider the fact that the prediction of the RUL of an equipment must consider its future dynamic usage and management, and the effects on its degradation. For example, the RUL of the gearbox of a wind turbine is influenced by the future loading conditions, which, in turn, depend on the wind conditions and on the O&M decisions that are taken for optimal equipment usage and for responding to power demand. In many prognostic systems, future conditions of equipment usage are generally assumed constant or behaving according to some known stochastic process, i.e., without considering the intertwined relation of RUL with O&M decisions (Ding *et al.*, 2018). Since this does not reflect reality, the RUL predictions that guide the O&M decisions are deemed to be incorrect and can lead to sub-optimal decisions (Bellani *et al.*, 2019). Also, O&M optimization of renewable energy systems should consider the availability of maintenance teams, the variability of demand and production, the long time horizons that characterize renewable energy usage and the uncertainty related to all the pieces of information.

In this context, the objective of the present work is to optimize O&M of renewable energy systems equipped with PHM capabilities. In order to deal with the issues presented above, the O&M management problem is formalized as a Sequential Decision Problem (SDP) over a long-time horizon. A SDP is characterized by the fact that the goodness of the selected action does not depend exclusively on the single decision, i.e. the goodness of the state entered as consequence of the selected action, but rather on the whole sequence of future decisions.

To solve the SDP, we adopt Deep Reinforcement Learning (DRL) (Sutton and Barto, 2018). Reinforcement Learning (RL) is a machine learning framework in which a learning agent optimizes its behavior by means of consecutive trial and error interactions with a white-box model of the system, i.e., a transparent and easily interpretable environment for the simulation of the system evolution, to find the optimal policy (Grondman *et al.*, 2012), i.e. the function linking each system state to the action that maximizes a reward. RL has been shown able to solve complex decision-making problems in many fields (Li, 2017), including energy-related ones (Rocchetta *et al.*, 2019).

Although, in principle, tabular RL algorithms allow finding the exact solution of SDPs, in most practical cases their computational cost is not compatible with applications to complex systems (Sutton and Barto, 2018; Tavares and Chaimowicz, 2018). For this reason, we resort to DRL, which uses Deep Neural Networks (DNNs) to find an

approximate solution of the optimization problem. In particular, we adopt the Proximal Policy Optimization (PPO) algorithm (Schulman *et al.*, 2017), which is one of the state-of-the-art approaches for DRL implementation.
The proposed framework is applied to a case study concerning the optimization of the O&M strategy of a wind farm. The application is meaningful since wind energy has become one of the most important alternatives for electricity production, with a growth rate larger than 10% in the last years according to the World Wind Energy Association (World Wind Energy Association, 2017). Furthermore, wind farms are characterized by O&M costs that can represent up to 20-25% of the entire life-cycle cost (Leite, Araújo and Rosas, 2018). For this reason, it is of utmost importance to develop methodologies to optimize O&M, to avoid unexpected outages due to failures and unnecessary maintenance interventions. The problem of maintenance optimization in wind farms has been reviewed in (Barberá *et al.*, 2013; Ding, Tian and Jin, 2013; Shafiee and Sørensen, 2019).
The main novelties of the proposed approach with respect to those already developed for O&M in wind farms are:

- the use of RUL predictions for O&M optimization;
- the fact of establishing the maintenance policy without any a-priori assumption on the type of maintenance strategy, e.g., corrective, scheduled, condition-based, predictive. This allows defining a completely assumptions-free approach for O&M optimization. Notice the improvement with respect to state-of-the-art works, which are limited to optimizing the parameters, e.g., maintenance period or degradation threshold, of an a-priori established maintenance strategy;
- the possibility of accounting for the influence of the dynamic environment and the effects of the O&M actions performed, on the future evolution of the system.

The effectiveness of the proposed approach is shown by means of a comparison with other state-of-the-art and user-defined O&M strategies, on a case study which considers a wind farm composed of 30 Wind Turbines (WTs).
The structure of the paper is as follows. In Section 2, we give an overview on state-of-the-art maintenance optimization methodologies. In Section 3, we introduce the problem statement and in Section 4 we discuss its formulation as a SDP. In Section 5, details about the RL algorithm adopted in this work are provided. In Section 6, the case study concerning the wind farm is presented. Results are discussed in Section 7. In Section 8, further experiments are proposed and analyzed, and conclusions are drawn in Section 9.

2. Maintenance in industrial systems

Many studies have shown the possibility of increasing production availability of industrial systems by improving the effectiveness of maintenance (Coit and Zio, 2019; de Jonge and Scarf, 2020), whose activities amount to one of the largest costs.
Maintenance optimization approaches have, thus, been developed. They can be classified according to different taxonomies: *i*) optimization algorithms, *ii*) optimization criteria, *iii*) outcomes of the optimization *iv*) characteristics of the system.
With respect to the type of optimization algorithm (taxonomy *i*), graphical methods (Labib and Yuniarto, 2009), Multiple Criteria Decision Making (MCDM) approaches based on Analytic Hierarchy Process (AHP) (Bevilacqua and Braglia, 2000), Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) (Ding and Kamaruddin, 2012), and Elimination Et Choice Translating Reality (ELECTRE) (Trojan and Morais, 2012), combination of grid search algorithms and simulation methods based on Monte Carlo simulation (Fedele and Zio, 2015; de Angelis, Patelli and Beer, 2017), Markov processes (Welte, Vatn and Heggset, 2006) and Petri nets (Santos, Teixeira and Soares, 2019), mixed integer programming (Nápoles-Rivera *et al.*, 2013), evolutionary algorithms (Haladuick and Dann, 2018; Mellal and Zio, 2019) and RL approaches (Kuhnle, Jakubik and Lanza, 2019; Rocchetta *et al.*, 2019) have been developed.

With respect to the optimization criteria (taxonomy *ii*), the most commonly used criteria are of economic and safety nature, such as maintenance cost (Lin, Li and Zio, 2018; Wang, Zhu and Yuan, 2018), life-cycle cost (Morcous and Lounis, 2005; Mellal, Zio and Williams, 2020), plant profit (Borgonovo, Marseguerra and Zio, 2000; Oke, 2005), availability (Laggoune, Ait Mokhtae and Kheloufi, 2011; Mellal and Zio, 2019), reliability (Marseguerra, Zio and Podofillini, 2004; Li, Guo and Zhou, 2016), resilience (Dehghani, Mohammadi Darestani and Shafieezadeh, 2020; Fang *et al.*, 2021). Both single objective and multi-objective approaches have been proposed. To address issues related to specific applications, other optimization criteria such as personnel management (Ni and Jin, 2012), spare parts inventory (Marseguerra, Zio and Podofillini, 2005; Ilgin and Tunali, 2007), environmental impact (García-Segura *et al.*, 2017) and production quality (Wang, Chu and Wu, 2007) have been considered.
With respect to the outcomes of the maintenance optimization (taxonomy *iii*), the methods can provide: *a*) indications to the technicians to assist in their maintenance decisions making and planning (Ben Said *et al.*, 2013), *b*) the best maintenance strategy among some a-priori defined alternatives (Haladuick and Dann, 2018), *c*) the optimal setting of the maintenance strategy, e.g., the optimal time interval between scheduled maintenance interventions (Compare, Martini and Zio, 2015; Javanmard and Koraeizadeh, 2016) or optimal degradation threshold in condition-based strategy (Marseguerra, Zio and Podofillini, 2002).
With respect to the characteristics of the considered systems (taxonomy *iv*), the methods can be distinguished between those addressing single-unit (Cha, Finkelstein and Levitin, 2017) and multi-unit systems (Vu *et al.*, 2014). Also systems characterized by different types of dependence among components have been considered: independent units (Bajestani and Banjevic, 2016), economic dependence, stochastic dependence, structural dependence and logistical dependence (Vu, Do and Barros, 2016; Farsi and Zio, 2020).

2.1. Maintenance in the wind power industry

For what concerns the optimization of maintenance in the wind power industry, various approaches have been proposed. In (Nielsen and Sørensen, 2011), a framework based on Bayesian updating is developed to optimize condition-based maintenance in offshore wind farms. In (Ding and Tian, 2011), an opportunistic maintenance plan has been optimized by simulating the effect of different parameters sets and considering the impact of imperfect maintenance. In (Tian *et al.*, 2011), a procedure to optimize failure probability thresholds assuming a condition-based maintenance strategy for WTs has been proposed. for In (Carlos et al., 2013), Genetic Algorithms (GAs) are used to optimize the scheduled maintenance strategy of a wind farm taking into account the stochasticity of wind power production. In (Nielsen and Sørensen, 2014), several methods for maintenance optimization of WTs, such as graphical, Bayesian and simulation-based approaches have been investigated. The authors have shown that the methods which make use of more sources of information and are able to provide time-variant policies are those which provide more satisfactory performance. In (Atashgar and Abdollahzadeh, 2016), an opportunistic maintenance strategy for a wind farm is optimized using particle swarm algorithm. In (Zhang *et al.*, 2017), the fruit fly optimization algorithm has been used to determine the optimal opportunistic maintenance threshold. In (Izquierdo et al., 2019), GAs are used to optimize an opportunistic maintenance strategy considering the dependencies among several components. In (Santos, Teixeira and Soares, 2019), a Petri net-based simulation approach is used to compare the performance of several maintenance strategies with respect to the minimization of the maintenance cost of a wind farm. The work has shown that opportunistic corrective maintenance allows obtaining the best performance in the considered case study. In (Zhou *et al.*, 2020), mixed integer linear programming has been used to discover cost-effective joint preventive maintenance plans for three wind farms. to In (Yang *et al.*, 2020), an opportunistic maintenance strategy for a wind farm is developed using an artificial bee colony algorithm, which considers information about wind and aging. In (Zhang and Yang, 2021), GAs are used to optimize the maintenance schedules of adjacent wind farms taking into account resource allocation.

All these literature works address the maintenance optimization problem by selecting state-of-the-art maintenance approaches and choosing the best performing one or by tuning the parameters of an a-priori selected maintenance approach, e.g., planned periodic or condition-based, to obtain the best possible result with respect to the selected optimization criteria. This implies that the search space is restricted to a limited number of state-of-the-art or user-defined maintenance strategies. Also, although many works have discussed the possibility of estimating the RUL of WTs (Ziegler *et al.*, 2018; Njiri *et al.*, 2019), according to the authors' best knowledge, no work has exploited this information in a maintenance optimization approach applied to wind farms. Finally, even if RL has been already applied to several maintenance optimization problems, its capability in dealing with maintenance optimization of renewable energy systems and, in particular, of wind farms, has not been discussed, yet.

3. Problem Statement

We consider a renewable energy system composed of L independently degrading units. The time horizon, T_M, is discretized into N_{T_M} decision times and we indicate the generic decision time as t. Maintenance is performed by a maintenance crew. At each decision time, the possible destinations of the maintenance crew are the L units or the depot, H. Once the maintenance crew reaches the generic $l-th$ unit, it performs: *i*) Preventive Maintenance (PM) if the unit is not failed, or *ii*) Corrective Maintenance (CM) if the unit is failed. Once the maintenance crew reaches the depot, H, it waits up to the next decision time, $t+1$. The downtimes of the units caused by PM and CM actions, Π_{PM} and Π_{CM}, are uncertain quantities, with the downtime of PM interventions expected to be smaller than that of CM interventions, as logistic support issues have already been addressed (Compare *et al.*, 2018).
The costs of preventive and corrective maintenance actions are U_{PM} and U_{CM}, respectively, and take into account the maintenance equipment costs and the maintenance crew costs.
The $l-th$ unit, $l \in \Lambda = \{1, \dots, L\}$, is equipped with a PHM system for the prediction of its RUL. A PHM system is typically composed of a monitoring system for the measurement, transmission and storing of the relevant physical quantities and algorithms for the evaluation of the system health state and prediction of the RUL (Aivaliotis, Georgoulias and Chryssolouris, 2018).
The production level $P_l(t)$ of the $l-th$ unit at time t represents the fraction of power produced at time t with respect to the absolute maximum power that can be produced by that unit. $P_l(t)$ depends on the environmental conditions, which are typically estimated in advance using data-driven approaches (Haddad *et al.*, 2019; Nazir *et al.*, 2020), and the component degradation state, which is related to the component age, Ag_l. We assume to have available a model predicting at any time t the present production level, $\hat{P}_l(t)$, and the future ones, $\hat{P}_l(t+j)$, for the following J days.
At any time t, the revenue generated from the total system production, $\sum_{l=1}^{L} P_l(t)$, is indicated as G_t.
The objective of the work is to define the optimal O&M policy, π^*, i.e., the optimal sequence of actions to be taken at every decision instant t in order to maximize the system profit, i.e., the difference between revenues and costs, over the time horizon T_M.

4. Problem Formulation

We formulate the problem as a SDP defined by the set $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma \rangle$, where:

- $\mathcal{S}$ is the state-space, i.e., the set of variables describing the state of the system;
- $\mathcal{A}$ is the action-space, i.e., the set of possible actions;
- $\mathcal{P}$ represents the transition probability, i.e., $\mathcal{P}(s'|s,a)$ is the probability of making a transition from state s to state s' by performing action a;

- $\mathcal{R}$ is the reward function, i.e., $\mathcal{R}(s'|s,a)$ is the reward which is received as results of reaching state s' after performing action a in state s, and it is used to update the policy;
- $\gamma \in [0,1]$ is the discount factor, i.e., the factor used to evaluate the present value of future rewards.

In Subsections 4.1, 4.2 and 4.3, the state-space, $\mathcal{S}$, the action-state, $\mathcal{A}$, and the reward function, $\mathcal{R}$, are defined, respectively. In Subsection 4.4, the developed model of the environment is described. Notice that, since in RL the learning agent directly interacts with the model of the environment, the explicit definition of the transition function $\mathcal{P}$ is not required.

4.1. State-space

The state at time t contains all the information retrievable from the energy renewable system and its environment. It is defined by the vector $\boldsymbol{s}_t = [\widehat{\boldsymbol{R}}_t^*, \widehat{\boldsymbol{P}}_t, \boldsymbol{MT}_t, \boldsymbol{Ag}_t, t]$, obtained appending the vectors of the units RULs predicted at time t by the PHM system, $\widehat{\boldsymbol{R}}_t^* = [\hat{R}_1^*(t), \dots, \hat{R}_L^*(t)]$, the predictions of the production of the units at time $t, t+1, \dots, t+J$, $\widehat{\boldsymbol{P}}_t = [\hat{P}_1(t), \dots \hat{P}_L(t), \hat{P}_1(t+1), \dots, \hat{P}_L(t+1), \dots, \hat{P}_1(t+J), \dots, \hat{P}_L(t+J)]$, the times needed to complete the ongoing maintenance actions, $\boldsymbol{MT}_t = [MT_1, \dots, MT_L]$, which are set to 0 if the units are not under maintenance at time t, the current ages of the units $\boldsymbol{Ag}_t = [Ag_1, \dots, Ag_L]$ and the current time t. The total dimensionality of the state-space is $(4+J) \cdot L + 1$.

4.2. Action-space

The possible destinations, i.e., the L units and the depot, are organized in the vector $\boldsymbol{A} = [a_1, \dots, a_{L+1}]$, where $a_l, l = 1, \dots, L$, refers to the $l-th$ unit and $L+1$ to the depot. At any time t, a decision is taken about the next destination of the maintenance crew. Namely, the learning agent returns as output a scalar $a_t \in \boldsymbol{A}$, that represents the destination of the crew. If one of the L units is selected as destination, the maintenance intervention, which can be preventive, if the unit is not failed, or corrective, if it is failed, starts as soon as the crew reaches the unit, whereas if the depot is selected as destination, the crew will start waiting for a new assignment as soon as it arrives at destination. When a maintenance operation starts, the corresponding unit is stopped and its production level becomes 0.

4.3. Reward function

At every decision instant t, the learning agent receives a reward r_t:

$$r_t = G_t - X_t \tag{2}$$

where the revenue G_t at time t is directly proportional to the total system production:

$$G_t = \sum_{l=1}^{L} K \cdot P_l(t) \tag{3}$$

being K the maximum revenue per unit, i.e., the revenue obtained when $P_l(t) = 1$. The maintenance cost X_t at time t is:

$$X_t = \sum_{l=1}^{L} U_{PM} \cdot I_{t<T_l}(t) \cdot I_{a_t=a_l}(t) + U_{CM} \cdot I_{t \geq T_l}(t) \cdot I_{a_t=a_l}(t) \tag{4}$$

where $I_{t<T_l}$, $I_{t\geq T_l}$ and $I_{a_t=a_l}$ are Boolean variables equal to 1 only when the condition at the subscript is satisfied. In practice, $I_{t\geq T_l}$ ($I_{t<T_l}$) indicates whether the component has (not) already failed at time t and therefore should

undertake corrective (preventive) maintenance. $I_{a_t=a_l}$ indicates whether the $l-th$ unit has been selected as destination for the maintenance crew at time t.

4.4. Model of the environment

Despite that the learning agent can discover the optimal O&M policy by means of direct interactions with the real-world system, this turns out to be unfeasible in the case of renewable energy systems for economic, safety and time issues (Sutton and Barto, 2018). Due to the trial-and-error nature of the learning process, the agent would need to perform several times actions suggested by the algorithm to explore their outcomes, leading to economically inconvenient and unsafe system management in the early stages of the learning process, when they are not yet optimal. Thus, the learning agent is trained using a white-box model of the system of interest.

The model of the environment developed in this work includes a stochastic model of the unit failure time, which is based on: *i*) a probability density function, $f_{T_l^*}(t)$, describing the failure time of the $l-th$ unit, T_l^*, assuming that it works until failure in nominal operating conditions *ii*) a law which allows computing the unit ground-truth failure time, T_l, considering T_l^* and the operating conditions actually experienced by the unit during its entire life. The white-box model of the system uses the two components in *i*) and *ii*) to represent the operating conditions' influence on the degradation process and consequent failure time.

At any decision time, $t \in \{1, \ldots, T_M\}$, the PHM system predicts the $l-th$ unit RUL, $\hat{R}_l^*(t),\ l = 1, \ldots, L$, assuming that it works in nominal operating conditions for the rest of its useful life. The prediction is affected by an error $\epsilon_R \sim N(0, \sigma_R)$, which describes the uncertainty due to the aleatory nature of the degradation process, the measurement error and the epistemic uncertainty of the prediction model (Baraldi, Mangili and Zio, 2013; Deng, Santos and Curran, 2020).

We assume to have available a model predicting at any time t the present production level, $\hat{P}_l(t)$, and the future ones, $\hat{P}_l(t+j), j = 1, \ldots, J,$ for the following J days,

$$\hat{P}_l(t+j) = P_l(t+j) + \epsilon_P \qquad j = 0, \ldots, J \tag{1}$$

where $P_l(t+j)$ is the ground-truth production level and $\epsilon_P \sim N(0, \sigma_P)$ is the model prediction error.

The training of the learning agent is performed using the white-box model of the environment, whereas the actual data collected from the real-world renewable energy system can be fed to the RL algorithm, which provides as output the O&M actions to be performed.

5. Reinforcement Learning Algorithms

A schematic view of the general RL procedure is shown in Figure 1. At each decision time, the learning agent observes the state of the environment and selects the action to be performed. This action leads to a change of the environment state and to a reward that is given as feedback to the agent for learning. By repeating this procedure several times, the learning agent discovers the optimal policy, π^*, which maps the possible environment states into the most suitable actions.

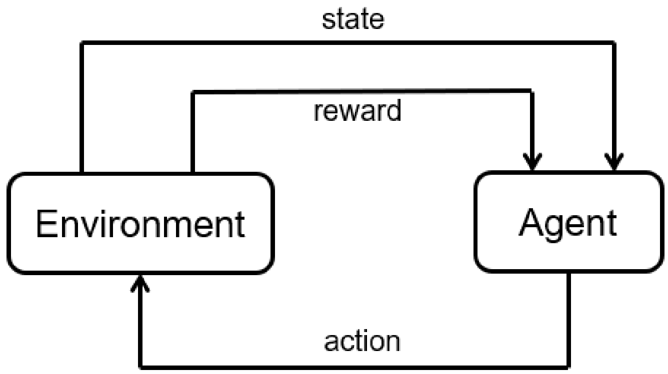


Figure 1. Schematic representation of RL.

RL algorithms can be classified into three groups: *policy search*, *value function* and *actor-critic* methods (Konda and Tsitsiklis, 2000). Policy search methods directly look for the optimal policy by learning a parameterized policy through which optimal actions are selected. The update of the policy parameters can be performed by means of gradient-free methods, e.g., evolutionary algorithms, or gradient-based methods, e.g. REINFORCE algorithms (Williams, 1992). Even if these methods have been shown to be effective in high dimensional or continuous actions spaces, they typically suffer from high variance in the estimates of the gradient and tend to converge to local optima rather than to the global optimum (Grondman *et al.*, 2012).

Differently, value function methods learn the value of being in a particular state and, then, select the optimal action according to the estimated state values. A well-known example of value function method is Deep Q-Networks (DQN) (Mnih *et al.*, 2015), in which a DNN is used to approximate the action-value function, $Q^{\pi}(s,a)$, for each possible state-action pair. Then, the optimal policy π^* is the one that maximizes the action-value function $Q^{\pi^*}(s,a)$:

$$\pi^* = argmax_a \, Q^{\pi}(s,a) \tag{5}$$

On one hand, the non-linear function approximation of the action-value function provided by the DNN allows dealing with complex systems for which an analytical treatment is unfeasible, but, on the other hand, it can introduce instability and divergence in the learning process, mainly because of two sources of correlations: among consecutive observations, and among the action-values and the target values of the learning process. Several improvements have been proposed to deal with this issue, such as experience replay and target networks (Mnih *et al.*, 2015). Experience replay stores transitions in a cyclic buffer from which training batches are randomly sampled in order to remove the correlations in the sequence of observations and to increase the method sample efficiency, whereas target networks rely on a second DNN, with a different set of weights from those used to select the most suitable action, to provide the target of the learning process. These weights are only periodically updated to remove the correlations between the action-value function Q and the target values. Value function methods usually show slow convergence rate and have been shown to fail on many simple problems (Schulman *et al.*, 2017).

Actor-Critic methods learn both the value function and the policy in an attempt to combine the strong points of value function and policy search methods (Konda and Tsitsiklis, 2000). Actor-Critic methods consist of two models: the critic, which learns the value function and the actor, which learns the policy by updating the parameters in the direction suggested by the critic.

In this work, the RL algorithm adopted to optimize O&M in a renewable energy system is PPO (Schulman *et al.*, 2017). PPO is an actor-critic algorithm, which aims at monotonically improving the policy during the learning process. PPO can be considered an enhancement of Trust Region Policy Optimization (TRPO) (Schulman *et al.*, 2015), in which the monotonicity of the improvement is guaranteed by means of a constraint that can be managed by means of second order approximations. The main idea is to avoid too large policy updates, which can increase the probability of accidental performance collapses. In PPO, the complexity of the second order approximations used in TRPO is overcome by clipping the objective function, which is defined as:

$$F = \mathbb{E}_t \left[\min \left(\frac{\pi(a|s)}{\pi_{old}(a|s)} \hat{A}^{\pi}(s,a), clip\left(\frac{\pi(a|s)}{\pi_{old}(a|s)}, 1-\epsilon, 1+\epsilon \right) \hat{A}^{\pi}(s,a) \right) \right] \tag{6}$$

where ϵ is an hyperparameter used to perform the clipping operation and $\hat{A}^{\pi}(s,a)$ is an estimator of the advantage function, defined as the difference between the action-value function, $Q^{\pi}(s,a)$, and the value function, $V^{\pi}(s)$, for a given state s:

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s) \tag{7}$$

The advantage function informs about the gain on the reward that can be obtained by performing a particular action a in state s, with respect to the reward obtained on average from that state. Its use allows reducing the variability of the objective function that would be obtained directly using the action-value function, $Q^{\pi}(s,a)$ (Baird III, 1993). According to Eq.(6), the objective function is defined as the minimum between an unclipped and a clipped version of the objective function used in TRPO (Schulman *et al.*, 2017). The minimum is used to define a lower, i.e., pessimistic, bound on the unclipped objective and the clipping operation is used as a regularizer that discourages to dramatically change the updated policy from the old one. PPO is considered relatively easy to implement and tune, and despite its simplicity, it has been shown able to outperform many state-of-the-art approaches on several benchmarks (Schulman *et al.*, 2017).

Finally, since the state space is very large, it can be hard for the agent to find the optimal action to be performed in every state in an efficient way starting from a random initialization of the neural network. This problem has been tackled by including domain knowledge in the learning process using methods such as reward shaping (Mataric, 1994) and state-action similarity solutions (Rosenfeld, Taylor and Kraus, 2017). In this work, we resort to Imitation Learning (IL) (Hester *et al.*, 2017), which consists in pre-training the agent to reproduce a heuristic policy by means of supervised learning and, then, fine-tuning the agent using RL. Notice that imitation learning allows exploiting the experts' knowledge about existing maintenance practices.

6. Case Study

We consider a wind farm composed of $L = 30$ identical 1.3MW WTs, each one equipped with a dedicated PHM system over a time horizon of $T_M = 5000$ days. We assume that a WT works in nominal conditions when its production level $P_l(t)$ is lower than 0.7. The failure time, T_l^*, of a WT operating in nominal conditions is distributed as an exponential distribution with failure rate $\lambda_f = 6.58 \cdot 10^{-3}$ days^{-1}, obtained by modeling the WT as a series equivalent of sub-systems, whose failure rates are set equal to the values reported in (Ozturk, Fthenakis and Faulstich, 2018). Sampled failure times lower than 75 days are not considered to assure an acceptable value of useful life after each maintenance intervention and to avoid the rise of behaviors associable to maintenance-induced failures, for which RUL after maintenance is lower than RUL before maintenance (Jackson and Mailler, 2013).

The effect of operation in non-nominal conditions characterized by $P_l(t) \geq 0.7$ is to increase the degradation speed and, therefore, to reduce the useful life of the WT (Figure 2). This is modeled by assuming that for each time step in which the WT operates at $P_l(t) \geq 0.7$, the useful life of the WT decreases of a random quantity $\nu \sim U(1{,}5)$ days. In practice, the ground-truth failure time, T_l, of the $l-th$ WT is given by:

$$T_l = T_l^* - \sum_{\tau=1}^{T_l} \nu_\tau \quad (8)$$

The PHM system of the $l-th$ WT provides at each time t a prediction $\hat{R}_l^*(t)$ of the WT RUL in nominal condition, $R_l^*(t)$. The RUL prediction is affected by a Gaussian error with mean equal to 0 and standard deviation $\sigma_R = 0.1 \cdot R_l^*(t)$. The variance of the error is decreasing as time passes in consideration of the fact that RUL predictions become more accurate as the WT approaches the failure time (Liu, Zio and Hu, 2018). The wind speed is simulated using the Markov model developed in (Shamshad et al., 2005). In particular, historical data are used to compute the transition probabilities of a Markov chain whose states represent different wind speed ranges. The Markov model is, then, used to generate wind speed trajectories of the desired length. In this work, we consider 33 wind velocity ranges of $1\frac{m}{s}$. Starting from the wind velocity trajectories, the power production is, then, estimated by means of the power curve shown in Figure 3, where $v_{cut-in} = 3.5\frac{m}{s}$, $v_{rated} = 13\frac{m}{s}$ and $v_{cut-out} = 25\frac{m}{s}$, according to the data available for 1.3MW WTs (Bauer and Matysik, 2011). Notice that the WT produces power only when the wind speed is in the range $[v_{cut-\mathrm{in}}, v_{cut-out}]$, since for values lower than v_{cut-in} the wind speed is too low for the turbine blades to start rotating and for values larger than $v_{cut-out}$ the WT is disconnected to avoid catastrophic failures. The nominal power value is reached for wind speed larger than or equal to v_{rated}.

The influence of the WT degradation on the power production is modeled assuming that the WT performance declines by 1.6% per year according to (Staffell and Green, 2014). This is implemented by accordingly reducing the maximum achievable power production at each time step. We consider both PM and CM to be perfect, i.e., the maximum achievable power production is restored to its original value after each maintenance intervention.

Figure 4 shows a simulated trajectory of the power produced by a WT.

We assume that there is a prediction algorithm that allows estimating the future power production for the following $J = 2$ days. Then, at every decision time t, the value of the predicted production, $\hat{P}_l$, for the present and following J days, is set according to Eq.(3) with $\sigma_P = 0.05$. The maintenance times are sampled from exponential distributions with repair rate $\mu_{PM} = 2.94\ \mathrm{days}^{-1}$ and $\mu_{CM} = 1.83\ \mathrm{days}^{-1}$, for preventive and corrective maintenance, respectively, setting μ_{PM} and μ_{CM} equal to the inverse of the mean values of the PM and CM repair times of different WT sub-systems (Carroll, McDonald and McMillan, 2016). The maximum daily revenue per unit is set equal to $K = 96$, whereas the cost of PM and CM actions are $U_{PM} = 180$ and $U_{CM} = 2247$ (Carroll, McDonald and McMillan, 2016), all in arbitrary units.

Finally, the discount factor, γ, as been set equal to 0.99.

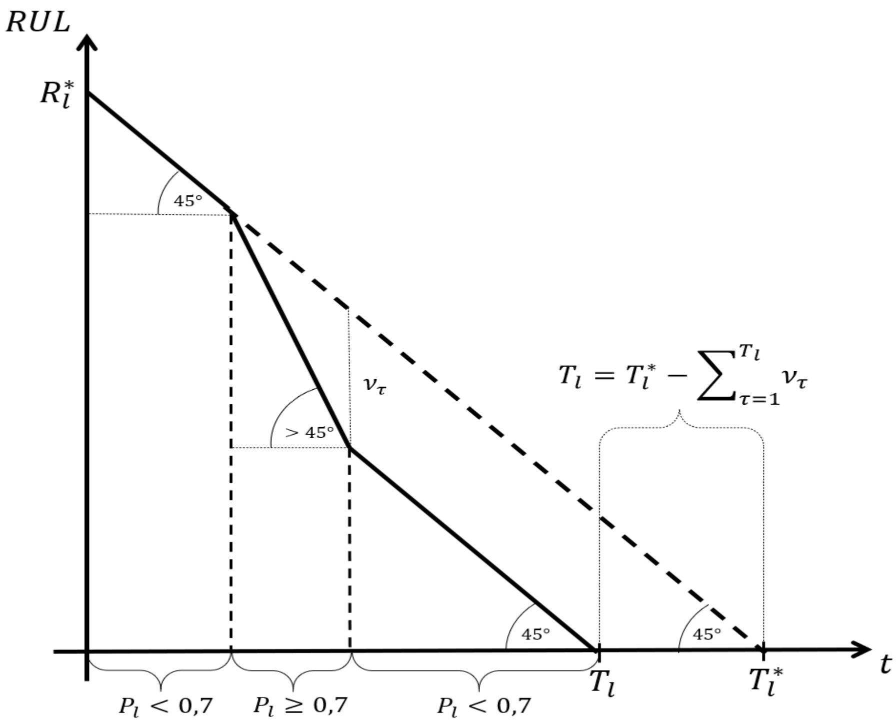


Figure 2. Dependence of the ground-truth failure time, T_l, on the operating conditions.

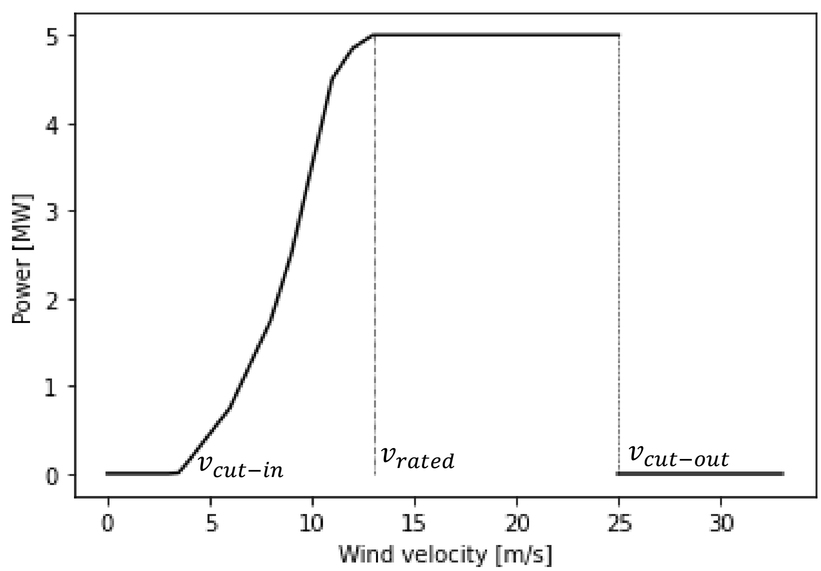


Figure 3. Wind turbine power curve.

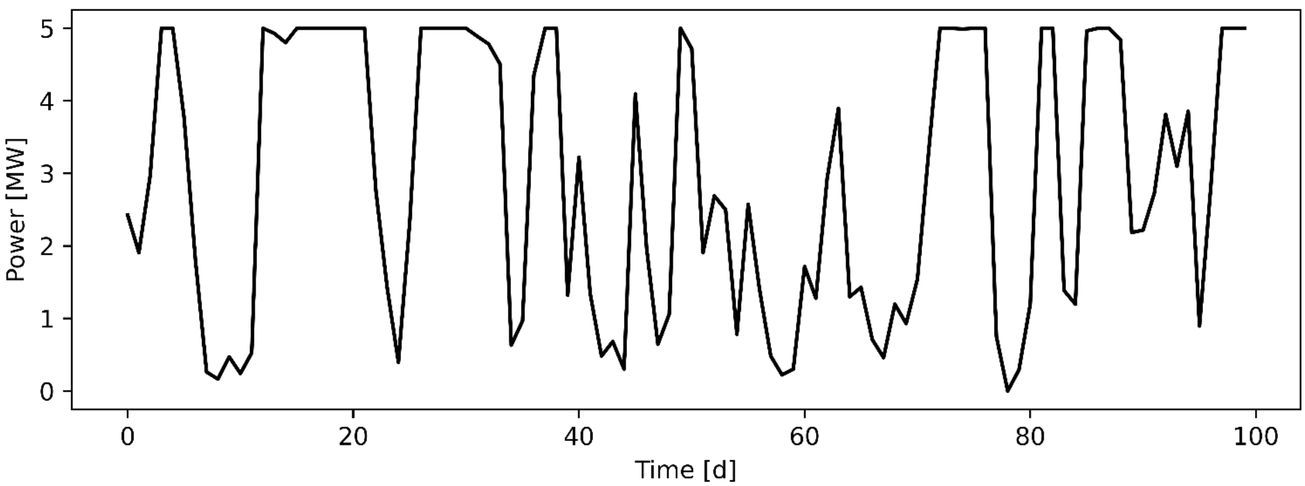


Figure 4. Example of simulated power production trajectory.

7. Results

We resort to a feedforward neural network characterized by 2 hidden layers of 64 neurons each, as learning agent. The IL step is performed by simulating 500 predictive maintenance trajectories and training the learning agent for 40 epochs. The PPO clipping hyperparameters ϵ is set equal to 0.2 and training lasts for a total of 10^6 time steps using 8 actors in parallel. The computations have been performed on two Intel® Xeon® CPUs at 2.30 GHz with 13 GB of RAM using Python.

The PPO-based RL optimized policy has been compared with the following user-defined strategies over 100 test episodes: *i*) a corrective maintenance strategy, *ii*) a scheduled maintenance strategy in which the maintenance interventions are scheduled at regular intervals, *iii*) a predictive maintenance strategy in which the maintenance interventions are performed only when the turbine RUL estimation is smaller than a user-defined threshold and i*v*) a predictive-heuristic maintenance strategy in which the maintenance intervention is planned when both the turbine RUL and future power production are below user-defined thresholds. The latter strategy has been introduced to consider the possibility of manually modifying the predictive maintenance strategy to take into account the information on the production level. The hyperparameters of all these maintenance strategies, i.e., time interval between two consecutive maintenance interventions for scheduled maintenance, degradation threshold for predictive maintenance, degradation and power thresholds for predictive-heuristic maintenance, have been set by optimizing the profit over 250 episodes using the Tree-structured Parzen Estimator (TPE) algorithm (Bergstra et al., 2011). The performance of the proposed PPO-based RL approach has been compared also to two other state-of-the-art RL algorithms: *v*) a value function method, i.e., DQN with experience replay and target network and *vi*) an actor-critic method, i.e., sample-efficient Actor-Critic with Experience Replay (ACER) (Wang *et al.*, 2017), which improves sample efficiency by introducing experience replay for actor-critic algorithms. The experience replay buffer size has been set equal to 50000 and to 5000 for DQN and ACER, respectively, and the training lasts $2 \cdot 10^6$ and 10^6 time steps, respectively, since DQN generally requires longer training times to converge.

The obtained performance over 100 test episodes are reported in Table 1. All the RL policies provide better performance than the corrective and scheduled maintenance strategies, which are the maintenance strategies most commonly applied to WTs (Pattison *et al.*, 2016). The DQN-based and the PPO-based policies are characterized by performances comparable to the predictive and predictive-heuristic strategies, which exploit the information about the equipment health state. The PPO-based policy is able to increase the profit of 1% with respect to the predictive strategy, despite that it does not reduce the number of preventive maintenance interventions and the number of failures. This is because the learning agent prefers to perform a larger number of maintenance interventions to keep the WT in low degradation states characterized by larger production levels and, on the other side, accepts the risk of failure when the predicted power production is large. Figure 5 shows the number of PM interventions performed by the predictive, predictive-heuristic and PPO-based policies, normalized by the number of time steps in which the unit production is at a given power level. The predictive strategy performs the same number of maintenance actions at every power level, the predictive-heuristic strategy performs many interventions at low power levels and few interventions at large power levels, with no interventions at power levels larger than 0.95, whereas the RL agent prefers to perform maintenance at low power levels but, differently from the predictive-heuristic strategy, it sometimes performs maintenance when the power is equal to one, in order to avoid failures. It is interesting to observe that, even if IL has been used to pre-train the RL agent to approximate the optimal predictive strategy, PPO is able to identify a different and better performing strategy.

Finally, considering the last column of Table 1, it can be noticed that the DRL-based approaches require larger computational times to identify the optimal policy than that required for the optimization of the maintenance interval (scheduled maintenance), RUL thresholds (predictive maintenance), RUL and power thresholds (predicitive-heuristic) by the TPE algorithm. This is due to the fact that the investigated RL approaches are composed of two stages (IL and RL) both requiring the training of a DNN, which is usually characterized by long computational time. Nevertheless, the computational times are still acceptable since they are limited to a few hours. Also, once the optimal policy has been found, it can be applied in almost real time to obtain the action to be performed given the environment data.

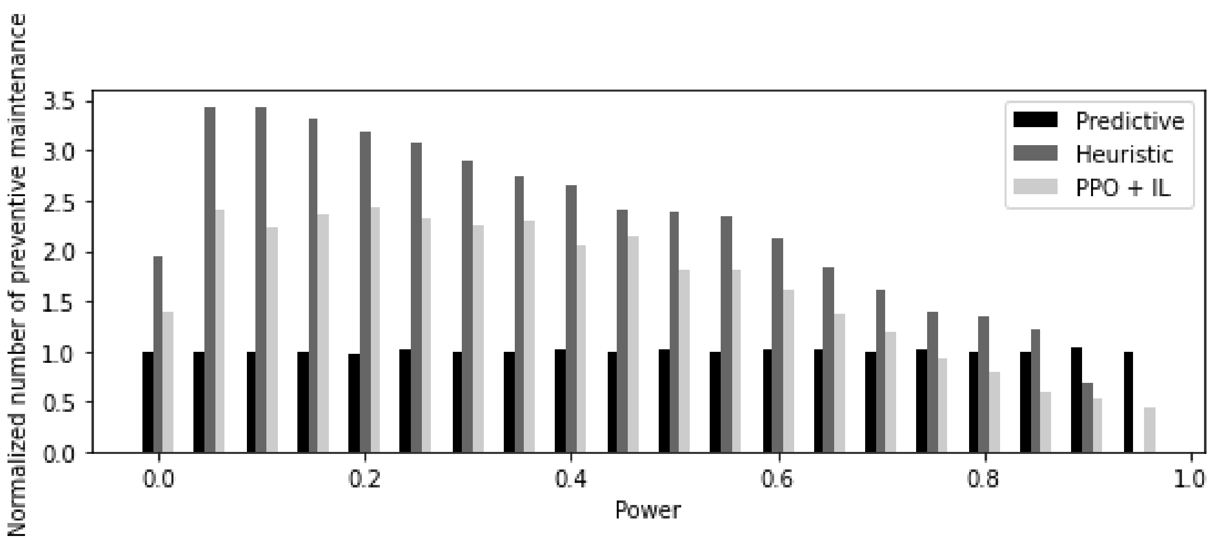


Figure 5. Normalized number of preventive maintenance interventions as a function of the power level.

Table 1. Performance of the tested strategies in terms of average profit over 100 test episodes.

Maintenance strategy	**Average profit (Ranking)**	**Number of corrective maintenance interventions (Ranking)**	**Number of preventive maintenance interventions (Ranking)**	**Computational time [s] (Ranking)**
Corrective	$(2.84 \pm 0.13) \cdot 10^6$ (7)	1550.91 ± 37.12 (7)	0.00 ± 0.00 (1)	0.00 (1)
Scheduled	$(4.10 \pm 0.08) \cdot 10^6$ (6)	1009.97 ± 25.54 (6)	2066.07 ± 25.82 (5)	2075.44 (4)
Predictive	$(7.27 \pm 0.01) \cdot 10^6$ (2)	0.51 ± 1.24 (1)	1685.15 ± 39.84 (3)	2031.95 (2)
Predictive-heuristic	$(7.25 \pm 0.02) \cdot 10^6$ (3)	1.55 ± 3.03 (2)	2070.44 ± 52.61 (6)	2070.35 (3)
PPO + IL	$(7.34 \pm 0.01) \cdot 10^6$ (1)	1.69 ± 1.46 (3)	1819.65 ± 45.42 (4)	11479.97 (5)
DQN + IL	$(6.69 \pm 0.11) \cdot 10^6$ (4)	43.53 ± 15.13 (4)	3139.84 ± 95.21 (7)	32148.44 (7)
ACER + IL	$(4.45 \pm 0.02) \cdot 10^6$ (5)	47.04 ± 4.67 (5)	1533.44 ± 49.17 (2)	16330.63 (6)

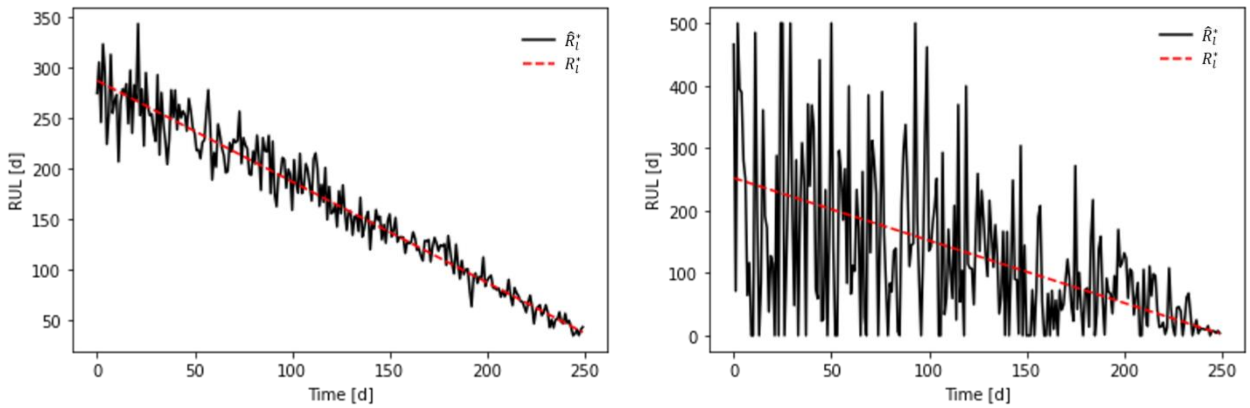
8. Further experiments

The robustness of the proposed method in discovering the optimal policy, π^*, is verified further by performing experiments which consider renewable energy systems with different characteristics and subject to unexpected issues.

8.1. Experiment 1

The objective of this experiment is to investigate the robustness of the proposed method with respect to possible underperformance of the PHM system, which can have several causes, such as sensor failures or the onset of degradation mechanisms not considered by the prognostic model. To this aim, we modify the case study presented in Section 6 by assuming that a subset of $L' = 5$ WTs, $\Lambda' = \{1, \dots, L'\} \subset \Lambda$, is equipped with PHM systems providing less accurate RUL predictions, specifically characterized by an error with a standard deviation $\sigma_R^{l'}$ equal to $0.99 \cdot R_{l'}^*$. Figure 6-a shows the accurate RUL predictions obtained for the generic $l-th$ WT, with $l \notin \Lambda'$, and

Figure 6-b shows the inaccurate RUL predictions of the $l' - th$ WT, with $l' \in \Lambda'$. The performance of the tested strategies over 100 test episodes are reported in Table 2.



(a) *(b)*

Figure 6. Comparison between the RUL prediction in case of small prediction error (a) and large prediction error (b).

The PPO-based RL policy is able to outperform the scheduled one, which is the best performing strategy among the state-of-the-art strategies and of the other RL approaches, providing an increment of 2.5% in terms of average profit with respect to the scheduled maintenance strategy.

Predictive and predictive-heuristic maintenance strategies perform a too large number of preventive maintenance interventions on the WTs equipped with underperforming PHM systems. PPO-based RL policy adopts a policy similar to a scheduled maintenance strategy for the five WTs of the set Λ', which allow reducing the number of unnecessary maintenance interventions without significantly increasing the number of corrective interventions. Finally, despite a suboptimal policy has been used in the IL step, i.e., the predictive maintenance strategy, the learning agent is able to find the best performing policy. This allows us to conclude that IL does not force the learning agent to converge to an a-priori selected maintenance policy and does not require the a-priori knowledge of the best performing maintenance policy.

Table 2. Comparison of the performance of the tested strategies in terms of average profit over 100 test episodes.

Maintenance strategy	**Average profit (Ranking)**	**Number of corrective maintenance interventions (Ranking)**	**Number of preventive maintenance interventions (Ranking)**	**Number of corrective maintenance interventions on units $l' \in \Lambda'$ (Ranking)**	**Number of preventive maintenance interventions on unit $l' \in \Lambda'$ (Ranking)**
Corrective	$(4.45 \pm 0.09) \cdot 10^6$ (6)	1385.86 ± 37.76 (7)	0.00 ± 0.00 (1)	233.43 ± 14.56 (7)	0.00 ± 0.00 (1)
Scheduled	$(6.81 \pm 0.02) \cdot 10^6$ (2)	40.84 ± 8.80 (2)	3606.06 ± 14.18 (4)	7.45 ± 3.52 (3)	599.71 ± 6.38 (2)
Predictive	$(6.67 \pm 0.02) \cdot 10^6$ (4)	14.93 ± 8.80 (1)	4422.08 ± 28.45 (7)	0.01 ± 0.10 (1)	3394.41 ± 27.26 (7)
Predictive-heuristic	$(6.70 \pm 0.04) \cdot 10^6$ (3)	52.37 ± 19.97 (4)	4032.24 ± 41.02 (5)	0.12 ± 0.40 (2)	2965.32 ± 39.83 (5)
PPO + IL	$(6.99 \pm 0.04) \cdot 10^6$ (1)	46.31 ± 10.27 (3)	2434.98 ± 40.64 (3)	14.45 ± 5.02 (5)	753.09 ± 23.30 (3)
DQN + IL	$(5.85 \pm 0.14) \cdot 10^6$ (5)	143.97 ± 19.81 (6)	4417.65 ± 51.11 (6)	19.47 ± 4.19 (6)	3347.79 ± 47.49 (6)
ACER + IL	$(4.25 \pm 0.08) \cdot 10^6$ (7)	61.66 ± 15.43 (5)	2073.90 ± 39.27 (2)	8.09 ± 5.66 (4)	1262.94 ± 29.24 (4)

8.2. Experiment 2

In this experiment, investigate whether the proposed approach is able to discover an optimal policy even in situations in which there is not a clear advantage in performing PM with respect to CM, and for which state-of-the-art and user-defined strategies are characterized by similar performance in terms of profit.

To this aim, the cost of PM has been increased to $U_{PM} = 1000$ arbitrary units, which is more than five times the cost considered in the case study of Section 6, the WT failure rate has been decreased to $\lambda_f = 1.81 \cdot 10^{-3}$ days^{-1}, which is less than one third of the failure rate considered in the case study of Section 6 and the WTs degrade their performance with a large degradation rate equal to 16% per year, which is ten times the degradation rate considered in the case study of Section 6. Also, the PM interventions are assumed to be imperfect, i.e., each PM intervention is characterized by a restoration factor q_{PM} sampled from a uniform distribution $U(0.35, 0.75)$. In practice, after each PM intervention, the age $Ag_l(t)$ of the $l-th$ unit is:

$$Ag_l(t) = Ag_l(t-1) - q_{PM} \cdot Ag_l(t-1) \tag{9}$$

In this experiment, CM is expected to perform better than PM in some circumstances, e.g., when a unit is very degraded (large age). Also, the impact of the age on the wind farm power production is amplified.

The performance of the tested strategies over 100 test episodes are reported in Table 3.

Table 3. Comparison of the performance of the tested strategies in terms of average profit over 100 test episodes.

Maintenance strategy	**Average profit (Ranking)**	**Number of corrective maintenance interventions (Ranking)**	**Number of preventive maintenance interventions (Ranking)**
Corrective	$(5.63 \pm 0.08) \cdot 10^6$ (6)	405.12 ± 23.03 (6)	0.00 ± 0.00 (1)
Scheduled	$(5.70 \pm 0.06) \cdot 10^6$ (5)	32.07 ± 20.63 (7)	396.16 ± 17.86 (3)
Predictive	$(5.72 \pm 0.12) \cdot 10^6$ (2)	0.00 ± 0.00 (1)	402.42 ± 25.35 (4)
Predictive-heuristic	$(5.71 \pm 0.11) \cdot 10^6$ (4)	0.61 ± 1.02 (2)	409.55 ± 29.45 (5)
PPO + IL	$(5.76 \pm 0.10) \cdot 10^6$ (1)	6.07 ± 2.85 (4)	426.38 ± 24.94 (6)
DQN + IL	$(5.71 \pm 0.12) \cdot 10^6$ (3)	3.53 ± 2.34 (3)	544.78 ± 37.54 (7)
ACER + IL	$(3.82 \pm 0.11) \cdot 10^6$ (7)	21.95 ± 3.93 (5)	272.8 ± 23.73 (2)

It can be noticed that all strategies are characterized by similar performances, except for the policy found by the ACER-based RL, that is not able to properly deal with the environment of this case study. In particular, the corrective and the scheduled maintenance strategies are now characterized by good performance since the CM are no more strongly penalized. The PPO-based RL policy is again the best performing policy with an increment of 0.7% in terms of average profit with respect to the predictive maintenance strategy. The PPO-based RL policy, differently from the predictive and the predictive-heuristic strategies and from the case study in Section 6, performs a larger number of CM interventions, which, however, allow increasing the profit since they improve the WTs health state more than the PM interventions.

9. Conclusions

A DRL-based approach for O&M optimization of renewable energy systems has been developed. It combines PPO, IL and a stochastic model of the environment which enables simulating the behavior of the renewable energy system. Its application to a wind farm of 30 WTs has shown that the proposed policy outperforms traditional maintenance strategies and other policies found by state-of-the-art DRL-based approaches, such as DQN and ACER, allowing increasing the average profit by 1% with respect to a predictive maintenance approach and by 10% with respect to DQN. Also, differently from the other approaches for maintenance optimization, which require to select a maintenance strategy and, then, optimize its parameters, the proposed approach does not require to select a-priori a maintenance strategy: it is able to automatically identify maintenance policies based on corrective or on preventive maintenance interventions, depending on the characteristics of the system, such as maintenance costs and accuracy of PHM algorithm predictions, and on the available sources of information and their uncertainties.

Future work will consider: *i*) the development of more advanced models of the environment which represent each unit as an engineering system formed by several interacting components, each one characterized by different degradation behavior, failure severity and impact on the power production, *ii*) the extension to the case in which more than one case maintenance crews are available and *iii*) the application of the O&M policy obtained using the model of the environment to data collected from a real-world renewable energy system.

References

Aivaliotis, P., Georgoulias, K. and Chryssolouris, G. (2018) 'A RUL calculation approach based on physical-based simulation models for predictive maintenance', *2017 International Conference on Engineering, Technology and Innovation: Engineering, Technology and Innovation Management Beyond 2020: New Challenges, New Approaches, ICE/ITMC 2017 - Proceedings*, 2018-Janua, pp. 1243–1246. doi: 10.1109/ICE.2017.8280022.

Al-Dulaimi, A. *et al.* (2019) 'A multimodal and hybrid deep neural network model for Remaining Useful Life estimation', *Computers in Industry*. Elsevier B.V., 108, pp. 186–196. doi: 10.1016/j.compind.2019.02.004.

de Angelis, M., Patelli, E. and Beer, M. (2017) 'Forced Monte Carlo Simulation Strategy for the Design of Maintenance Plans with Multiple Inspections', *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 3(2), pp. 1–9. doi: 10.1061/ajrua6.0000868.

Atashgar, K. and Abdollahzadeh, H. (2016) 'Reliability optimization of wind farms considering redundancy and opportunistic maintenance strategy', *Energy Conversion and Management*. Elsevier Ltd, 112, pp. 445–458. doi: 10.1016/j.enconman.2016.01.027.

Baird III, L. C. (1993) *Advantage updating*.

Bajestani, M. A. and Banjevic, D. (2016) 'Calendar-based age replacement policy with dependent renewal cycles', *IIE Transactions (Institute of Industrial Engineers)*. Taylor & Francis, 48(11), pp. 1016–1026. doi: 10.1080/0740817X.2016.1163444.

Baraldi, P., Mangili, F. and Zio, E. (2013) 'Investigation of uncertainty treatment capability of model-based and data-driven prognostic methods using simulated data', *Reliability Engineering and System Safety*. Elsevier, 112, pp. 94–108. doi: 10.1016/j.ress.2012.12.004.

Barberá, L. *et al.* (2013) 'State of the art of maintenance applied to wind turbines', *Chemical Engineering Transactions*, 33(January), pp. 931–936. doi: 10.3303/CET1333156.

Barreto, L., Amaral, A. and Pereira, T. (2017) 'Industry 4.0 implications in logistics: an overview', *Procedia Manufacturing*. Elsevier B.V., 13, pp. 1245–1252. doi: 10.1016/j.promfg.2017.09.045.

Bauer, L. and Matysik, S. (2011) *wind-turbine-models.com, Accessed on November 2020*. Available at: https://en.wind-turbine-models.com/.

Bellani, L. *et al.* (2019) 'Towards Developing a Novel Framework for Practical PHM: a Sequential Decision Problem solved by Reinforcement Learning and Artificial Neural Networks', *International Journal of Prognostics and Health Management*, 31, pp. 1–15. Available at: https://www.researchgate.net/publication/339016560.

Bergstra, J. *et al.* (2011) 'Algorithms for hyper-parameter optimization', *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, NIPS 2011*, pp. 1–9.

Bevilacqua, M. and Braglia, M. (2000) 'Analytic hierarchy process applied to maintenance strategy selection', *Reliability Engineering and System Safety*, 70(1), pp. 71–83. doi: 10.1016/S0951-8320(00)00047-8.

Borgonovo, E., Marseguerra, M. and Zio, E. (2000) 'A Monte Carlo methodological approach to plant availability modeling with maintenance, aging and obsolescence', *Reliability Engineering and System Safety*, 67(1), pp. 61–73. doi: 10.1016/S0951-8320(99)00046-0.

Cai, B. *et al.* (2020) 'Remaining Useful Life Estimation of Structure Systems under the Influence of Multiple Causes: Subsea Pipelines as a Case Study', *IEEE Transactions on Industrial Electronics*. IEEE, 67(7), pp. 5737–5747. doi: 10.1109/TIE.2019.2931491.

Carlos, S. *et al.* (2013) 'Onshore wind farms maintenance optimization using a stochastic model', *Mathematical and Computer Modelling*. Elsevier Ltd, 57(7–8), pp. 1884–1890. doi: 10.1016/j.mcm.2011.12.025.

Carroll, J., McDonald, A. and McMillan, D. (2016) 'Failure rate, repair time and unscheduled O&M costanalysis of offshore wind turbines', *Wind Energy*, 19, pp. 1107–1119. doi: 10.1002/we.

Cha, J. H., Finkelstein, M. and Levitin, G. (2017) 'On preventive maintenance of systems with lifetimes dependent on a random shock process', *Reliability Engineering and System Safety*. Elsevier Ltd, 168(March), pp. 90–97. doi: 10.1016/j.ress.2017.03.023.

Coit, D. W. and Zio, E. (2019) 'The evolution of system reliability optimization', *Reliability Engineering and System Safety*. Elsevier Ltd, 192(May 2018), p. 106259. doi: 10.1016/j.ress.2018.09.008.

Compare, M. *et al.* (2018) 'Reinforcement learning-based flow management of gas turbine parts under stochastic failures', *International Journal of Advanced Manufacturing Technology, Springer Verlag*. The International Journal of Advanced Manufacturing Technology, Springer Verlag, 99(9–12), pp. 2981–2992.

Compare, M., Baraldi, P. and Zio, E. (2020) 'Challenges to IoT-Enabled Predictive Maintenance for Industry 4.0', *IEEE Internet of Things Journal*. IEEE, 7(5), pp. 4585–4597. doi: 10.1109/JIOT.2019.2957029.

Compare, M., Martini, F. and Zio, E. (2015) 'Genetic algorithms for condition-based maintenance optimization under uncertainty', *European Journal of Operational Research*, 244(2), pp. 611–623. doi: 10.1016/j.ejor.2015.01.057.

Dehghani, N. L., Mohammadi Darestani, Y. and Shafieezadeh, A. (2020) 'Optimal life-cycle resilience enhancement of aging power distribution systems: A MINLP-Based preventive maintenance planning', *IEEE Access*, 8, pp. 22324–22334. doi: 10.1109/ACCESS.2020.2969997.

Deng, Q., Santos, B. F. and Curran, R. (2020) 'A practical dynamic programming based methodology for aircraft maintenance check scheduling optimization', *European Journal of Operational Research*. Elsevier B.V., 281(2), pp. 256–273. doi: 10.1016/j.ejor.2019.08.025.

Ding, F. *et al.* (2018) 'An integrated approach for wind turbine gearbox fatigue life prediction considering instantaneously varying load conditions', *Renewable Energy*. Elsevier Ltd, 129, pp. 260–270. doi: 10.1016/j.renene.2018.05.074.

Ding, F. and Tian, Z. (2011) 'Opportunistic maintenance optimization for wind turbine systems considering imperfect maintenance actions', *International Journal of Reliability, Quality and Safety Engineering*, 18(5), pp. 463–481. doi: 10.1142/S0218539311004196.

Ding, F., Tian, Z. and Jin, T. (2013) 'Maintenance modeling and optimization for wind turbine systems: A review', *QR2MSE 2013 - Proceedings of 2013 International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering*, pp. 569–575. doi: 10.1109/QR2MSE.2013.6625648.

Ding, S.-H. and Kamaruddin, S. (2012) 'Selection of Optimal Maintenance Policy by Using Fuzzy Multi Criteria Decision Making Method', in *Proceedings of the 2012 International Conference on Industrial Engineering and Operations Management*, pp. 435–443.

Fang, Y. *et al.* (2021) 'Resilient Critical Infrastructure Planning Under Disruptions Considering Recovery Scheduling', *IEEE Transactions on Engineering Management*. IEEE, 68(2), pp. 1–15.

Farsi, M. A. and Zio, E. (2020) 'Modeling and Analyzing Supporting Systems for Smart Manufacturing Systems with Stochastic, Technical and Economic Dependences', *International Journal of Engineering, Transactions B: Applications*, 33(11), pp. 2310–2318. doi: 10.5829/ije.2020.33.11b.21.

Fedele, L. and Zio, E. (2015) 'An Innovative Methodology for Cost Optimization of the Maintenance of Medical Devices', *IFMBE Proceedings*, 45, pp. 637–640. doi: 10.1007/978-3-319-11128-5_159.

García-Segura, T. *et al.* (2017) 'Lifetime reliability-based optimization of post-tensioned box-girder bridges', *Engineering Structures*, 145, pp. 381–391. doi: 10.1016/j.engstruct.2017.05.013.

Grondman, I. *et al.* (2012) 'A survey of actor-critic reinforcement learning: Standard and natural policy gradients', *IEEE Transactions on Systems, Man and Cybernetics Part C: Applications and Reviews*, 42(6), pp. 1291–1307. doi: 10.1109/TSMCC.2012.2218595.

Haddad, M. *et al.* (2019) 'Wind and Solar Forecasting for Renewable Energy System using SARIMA-based Model', in *6th International conference on Time Series and Forecasting*.

Haladuick, S. and Dann, M. R. (2018) 'Genetic algorithm for inspection and maintenance planning of deteriorating structural systems: Application to pressure vessels', *Infrastructures*, 3(3). doi: 10.3390/infrastructures3030032.

Hester, T. *et al.* (2017) 'Deep q-learning from demonstrations', *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, pp. 3223–3230.

Ilgin, M. A. and Tunali, S. (2007) 'Joint optimization of spare parts inventory and maintenance policies using genetic algorithms', *International Journal of Advanced Manufacturing Technology*, 34(5–6), pp. 594–604. doi: 10.1007/s00170-006-0618-z.

Izquierdo, J. *et al.* (2019) 'Framework for managing the operations and maintenance of wind farms', *Journal of Physics: Conference Series*, 1222(1). doi: 10.1088/1742-6596/1222/1/012046.

Jackson, C. and Mailler, B. (2013) 'Post-servicing failure rates: Optimizing preventive maintenance interval and quantifying maintenance induced failure in repairable systems', *Proceedings - Annual Reliability and Maintainability Symposium*, (January 2013). doi: 10.1109/RAMS.2013.6517681.

Javanmard, H. and Koraeizadeh, A. al W. (2016) 'Optimizing the preventive maintenance scheduling by genetic algorithm based on cost and reliability in National Iranian Drilling Company', *Journal of Industrial Engineering International*. Springer Berlin Heidelberg, 12(4), pp. 509–516. doi: 10.1007/s40092-016-0155-9.

de Jonge, B. and Scarf, P. A. (2020) 'A review on maintenance optimization', *European Journal of Operational Research*. Elsevier B.V., 285(3), pp. 805–824. doi: 10.1016/j.ejor.2019.09.047.

Konda, V. R. and Tsitsiklis, J. N. (2000) 'Actor-critic algorithms', *Advances in Neural Information Processing Systems*, pp. 1008–1014.

Kuhnle, A., Jakubik, J. and Lanza, G. (2019) 'Reinforcement learning for opportunistic maintenance optimization', *Production Engineering*. Springer Berlin Heidelberg, 13(1), pp. 33–41. doi: 10.1007/s11740-018-0855-7.

Kwon, D. *et al.* (2016) 'IoT-Based Prognostics and Systems Health Management for Industrial Applications', *IEEE Access*, 4, pp. 3659–3670. doi: 10.1109/ACCESS.2016.2587754.

Labib, A. and Yuniarto, M. N. (2009) 'Maintenance strategies for changeable manufacturing', in *Changeable and reconfigurable manufacturing systems*. Springer S. London: Springer, pp. 327–351. doi: 10.1007/978-1-84882-067-8.

Laggoune, R., Ait Mokhtae, W. and Kheloufi, K. (2011) 'Preventive maintenance optimization based on both cost and availability measures. A case study', in *ESReDA Conference 2011 Advances in Reliability-based Maintenance Policies*.

Leite, G. de N. P., Araújo, A. M. and Rosas, P. A. C. (2018) 'Prognostic techniques applied to maintenance of wind turbines: a concise and specific review', *Renewable and Sustainable Energy Reviews*, 81(February), pp. 1917–1925. doi: 10.1016/j.rser.2017.06.002.

Li, Y. (2017) 'Deep Reinforcement Learning: An Overview', *arXiv preprint arXiv:1701.07274*. doi: 10.1007/978-3-319-56991-8_32.

Li, Z., Guo, J. and Zhou, R. (2016) 'Maintenance scheduling optimization based on reliability and prognostics information', *Proceedings - Annual Reliability and Maintainability Symposium*, 2016-April(November 2017). doi: 10.1109/RAMS.2016.7448069.

Lin, Y. H., Li, Y. F. and Zio, E. (2018) 'A Framework for Modeling and Optimizing Maintenance in Systems Considering Epistemic Uncertainty and Degradation Dependence Based on PDMPs', *IEEE Transactions on Industrial Informatics*. IEEE, 14(1), pp. 210–220. doi: 10.1109/TII.2017.2743820.

Liu, J., Zio, E. and Hu, Y. (2018) 'Particle Filtering for Prognostics of a Newly Designed Product With a New Parameters Initialization Strategy Based on Reliability Test Data', *IEEE Access*. IEEE, 6, pp. 62564–62573. doi: 10.1109/ACCESS.2018.2876457.

Marseguerra, M., Zio, E. and Podofillini, L. (2002) 'Condition-based maintenance optimization by means of genetic algorithms and Monte Carlo simulation', *Reliability Engineering and System Safety*, 77(2), pp. 151–165. doi: 10.1016/S0951-8320(02)00043-1.

Marseguerra, M., Zio, E. and Podofillini, L. (2004) 'Optimal reliability/availability of uncertain systems via multi-objective genetic algorithms', *IEEE Transactions on Reliability*, 53(3), pp. 424–434. doi: 10.1109/TR.2004.833318.

Marseguerra, M., Zio, E. and Podofillini, L. (2005) 'Multiobjective spare part allocation by means of genetic algorithms and

Monte Carlo simulation', *Reliability Engineering and System Safety*, 87(3), pp. 325–335. doi: 10.1016/j.ress.2004.06.002.

Mataric, M. J. (1994) 'Reward Functions for Accelerated Learning', in *Machine Learning Proceedings*. Morgan Kaufmann Publishers, Inc., pp. 181–189. doi: 10.1016/b978-1-55860-335-6.50030-1.

Mellal, M. A. and Zio, E. (2019) 'Availability Optimization of Parallel-Series System by Evolutionary Computation', *Proceedings - 2018 3rd International Conference on System Reliability and Safety, ICSRS 2018*. IEEE, (3), pp. 198–202. doi: 10.1109/ICSRS.2018.8688722.

Mellal, M. A., Zio, E. and Williams, E. J. (2020) 'Cost minimization of repairable systems subject to availability constraints using efficient cuckoo optimization algorithm', *Quality and Reliability Engineering International*, 36(3), pp. 1098–1110. doi: 10.1002/qre.2617.

Mnih, V. *et al.* (2015) 'Human-level control through deep reinforcement learning', *Nature*. Nature Publishing Group, 518(7540), pp. 529–533. doi: 10.1038/nature14236.

Morcous, G. and Lounis, Z. (2005) 'Maintenance optimization of infrastructure networks using genetic algorithms', *Automation in Construction*, 14(1), pp. 129–142. doi: 10.1016/j.autcon.2004.08.014.

Nápoles-Rivera, F. *et al.* (2013) 'Simultaneous optimization of energy management, biocide dosing and maintenance scheduling of thermally integrated facilities', *Energy Conversion and Management*, 68, pp. 177–192. doi: 10.1016/j.enconman.2012.09.033.

Nazir, M. S. *et al.* (2020) 'Wind generation forecasting methods and proliferation of artificial neural network: A review of five years research trend', *Sustainability*, 12(9). doi: 10.3390/su12093778.

Ni, J. and Jin, X. (2012) 'Decision support systems for effective maintenance operations', *CIRP Annals - Manufacturing Technology*. CIRP, 61(1), pp. 411–414. doi: 10.1016/j.cirp.2012.03.065.

Nielsen, J. J. and Sørensen, J. D. (2011) 'On risk-based operation and maintenance of offshore wind turbine components', *Reliability Engineering and System Safety*. Elsevier, 96(1), pp. 218–229. doi: 10.1016/j.ress.2010.07.007.

Nielsen, J. S. and Sørensen, J. D. (2014) 'Methods for Risk-Based Planning of O&M of Wind Turbines', *Energies*, 7(10), pp. 6645–6664. doi: 10.3390/en7106645.

Njiri, J. G. *et al.* (2019) 'Consideration of lifetime and fatigue load in wind turbine control', *Renewable Energy*, 131, pp. 818–828. doi: 10.1016/j.renene.2018.07.109.

Oke, S. A. (2005) 'An analytical model for the optimisation of maintenance profitability', *International Journal of Productivity and Performance Management*, 54(2), pp. 113–136. doi: 10.1108/17410400510576612.

Okoh, C. *et al.* (2014) 'Overview of Remaining Useful Life prediction techniques in Through-life Engineering Services', *Procedia CIRP*. Elsevier B.V., 16, pp. 158–163. doi: 10.1016/j.procir.2014.02.006.

Ozturk, S., Fthenakis, V. and Faulstich, S. (2018) 'Failure modes, effects and criticality analysis for wind turbines considering climatic regions and comparing geared and direct drive wind turbines', *Energies*, 11(9). doi: 10.3390/en11092317.

Pattison, D. *et al.* (2016) 'Intelligent integrated maintenance for wind powergeneration', *Wind Energy*, 19(May 2015), pp. 547–562. doi: 10.1002/we.

Rocchetta, R. *et al.* (2019) 'A reinforcement learning framework for optimal operation and maintenance of power grids', *Applied Energy*, (May), pp. 291–301. doi: 10.1016/j.apenergy.2019.03.027.

Rosenfeld, A., Taylor, M. E. and Kraus, S. (2017) 'Leveraging human knowledge in tabular reinforcement learning: A study of human subjects', in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17)*, pp. 3823–3830. doi: 10.1017/S0000000000000000.

Ben Said, A. *et al.* (2013) 'A Bayesian network based approach to improve the effectiveness of maintenance actions in Semiconductor Industry', *A Bayesian Network based approach to improve the effectiveness of maintenance actions in Semiconductor industry*, pp. 1–11. Available at: http://www.phmsociety.org/sites/phmsociety.org/files/phm_submission/2014/phmce_14_024.pdf%0Ahttps://www.phmsociety.org/sites/phmsociety.org/files/phm_submission/2014/phmce_14_024.pdf.

Santos, F. P., Teixeira, Â. P. and Soares, C. G. (2019) 'Modeling, simulation and optimization of maintenance cost aspects on multi-unit systems by stochastic Petri nets with predicates', *Simulation*, 95(5), pp. 461–478. doi: 10.1177/0037549718782655.

Sanz-Bobi, M. A. (2014) *Use, Operation and Maintenance of Renewable Energy Systems: Experiences and Future Approaches*. Springer. Edited by Springer. Springer.

Schulman, J. *et al.* (2015) 'Trust region policy optimization', *32nd International Conference on Machine Learning, ICML 2015*, 3, pp. 1889–1897.

Schulman, J. *et al.* (2017) 'Proximal Policy Optimization Algorithms', *arXiv 2017." arXiv preprint arXiv:1707.06347.*, pp.

1–12.

Shafiee, M. and Sørensen, J. D. (2019) ‘Maintenance optimization and inspection planning of wind energy assets: Models, methods and strategies’, *Reliability Engineering and System Safety*. Elsevier Ltd, 192(November 2017). doi: 10.1016/j.ress.2017.10.025.

Shamshad, A. *et al.* (2005) ‘First and second order Markov chain models for synthetic generation of wind speed time series’, *Energy*, 30(5), pp. 693–708. doi: 10.1016/j.energy.2004.05.026.

Simões, J. M., Gomes, C. F. and Yasin, M. M. (2011) ‘A literature review of maintenance performance measurement: A conceptual framework and directions for future research’, *Journal of Quality in Maintenance Engineering*, 17(2), pp. 116–137. doi: 10.1108/13552511111134565.

Staffell, I. and Green, R. (2014) ‘How does wind farm performance decline with age?’, *Renewable Energy*. Elsevier Ltd, 66, pp. 775–786. doi: 10.1016/j.renene.2013.10.041.

Sutton, R. S. and Barto, A. (2018) *Reinforcement Learning: An Introduction*.

Tavares, A. R. and Chaimowicz, L. (2018) ‘Tabular Reinforcement Learning in Real-Time Strategy Games via Options’, *IEEE Conference on Computatonal Intelligence and Games, CIG*. IEEE, 2018-Augus. doi: 10.1109/CIG.2018.8490427.

Terrissa, L. S. *et al.* (2016) ‘A new approach of PHM as a service in cloud computing’, *Colloquium in Information Science and Technology, CIST*, 0, pp. 610–614. doi: 10.1109/CIST.2016.7804958.

Tian, Z. *et al.* (2011) ‘Condition based maintenance optimization for wind power generation systems under continuous monitoring’, *Renewable Energy*. Elsevier Ltd, 36(5), pp. 1502–1509. doi: 10.1016/j.renene.2010.10.028.

Tjahjono, B. *et al.* (2017) ‘What does Industry 4.0 mean to Supply Chain?’, *Procedia Manufacturing*. Elsevier B.V., 13, pp. 1175–1182. doi: 10.1016/j.promfg.2017.09.191.

Trojan, F. and Morais, D. C. (2012) ‘Prioritising alternatives for maintenance of water distribution networks : A group decision approach’, *Water SA*, 38(4), pp. 555–564.

Ustundag, A. and Cevikcan, E. (2018) *Industry 4.0: Managing The Digital Transformation*. Springer S. Edited by Spinger. Springer. doi: 10.1007/978-3-319-57870-5_7.

Vu, H. C. *et al.* (2014) ‘Maintenance grouping strategy for multi-component systems with dynamic contexts’, *Reliability Engineering and System Safety*. Elsevier, 132, pp. 233–249. doi: 10.1016/j.ress.2014.08.002.

Vu, H. C., Do, P. and Barros, A. (2016) ‘Mean Residual Life and the Birnbaum Importance Measure for Complex Structures’, 65(1), pp. 217–234.

Wang, J., Zhu, X. and Yuan, T. (2018) ‘Cost-Minimization Preventive Maintenance for the Data Storage System of a Supercomputer’, *Proceedings - 12th International Conference on Reliability, Maintainability, and Safety, ICRMS 2018*. IEEE, pp. 448–451. doi: 10.1109/ICRMS.2018.00089.

Wang, L., Chu, J. and Wu, J. (2007) ‘Selection of optimum maintenance strategies based on a fuzzy analytic hierarchy process’, *International Journal of Production Economics*, 107(1), pp. 151–163. doi: 10.1016/j.ijpe.2006.08.005.

Wang, Z. *et al.* (2017) ‘Sample efficient actor-critic with experience replay’, *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, (2016).

Welte, T. M., Vatn, J. and Heggset, J. (2006) ‘Markov state model for optimization of maintenance and renewal of hydro power components’, *2006 9th International Conference on Probabilistic Methods Applied to Power Systems, PMAPS*, (July 2006). doi: 10.1109/PMAPS.2006.360311.

Williams, R. J. (1992) ‘Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning’, *Machine Learning*, 8(3), pp. 229–256. doi: 10.1023/A:1022672621406.

World Wind Energy Association, (WWEA) (2017) *World Wind Market has reached 486 GW from where 54 GW has been installed last year*. Available at: https://wwindea.org/blog/2017/06/08/11961-2/.

Yang, L. *et al.* (2020) ‘Operations & Maintenance Optimization of Wind Turbines Integrating Wind and Aging Information’, *IEEE Transactions on Sustainable Energy*, 3029(c), pp. 1–1. doi: 10.1109/tste.2020.2986586.

Zhang, C. *et al.* (2017) ‘Opportunistic maintenance for wind turbines considering imperfect, reliability-based maintenance’, *Renewable Energy*. Elsevier Ltd, 103, pp. 606–612. doi: 10.1016/j.renene.2016.10.072.

Zhang, C. and Yang, T. (2021) ‘Optimal maintenance planning and resource allocation for wind farms based on non-dominated sorting genetic algorithm-II’, *Renewable Energy*. Elsevier Ltd, 164, pp. 1540–1549. doi: 10.1016/j.renene.2020.10.125.

Zhou, Y. *et al.* (2020) ‘Bio-objective long-term maintenance scheduling for wind turbines in multiple wind farms’, *Renewable Energy*. Elsevier Ltd, 160, pp. 1136–1147. doi: 10.1016/j.renene.2020.07.065.

Ziegler, L. *et al.* (2018) ‘Lifetime extension of onshore wind turbines : A review covering Germany, Spain, Denmark, and the UK’, *Renewable and Sustainable Energy Reviews*. Elsevier Ltd, 82(January 2017), pp. 1261–1271. doi: 10.1016/j.rser.2017.09.100.