Applied Energy 219 (2018) 53–67

Contents lists available at ScienceDirect

Applied Energy

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

Fuzzy Q-Learning for multi-agent decentralized energy management in microgrids

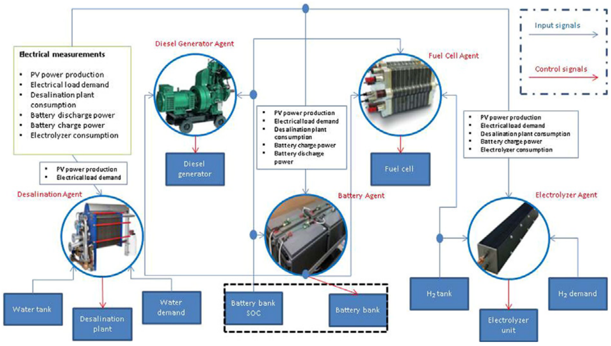
P. Kofinas[a,b,*], A.I. Dounis[b], G.A. Vouros[a]

[a] Department of Digital Systems, University of Piraeus, Piraeus, Greece
[b] Department of Industrial Design and Production Engineering, University of West Attica, Egaleo-Athens, Greece

HIGHLIGHTS

- Power balancing with a fully decentralized framework.
- MAS with modified Independent Learners approach for energy management of microgrid.
- MAS and Fuzzy Q-Learning for continuous states and actions space.
- Reinforcement Learning (Q-learning) for Collaborative MAS.

GRAPHICAL ABSTRACT



ARTICLE INFO

Keywords:
Energy management
Reinforcement learning (RL)
Fuzzy Q-Learning
Multi-agent system (MAS)
Microgrid

ABSTRACT

This study proposes a cooperative multi-agent system for managing the energy of a stand-alone microgrid. The multi-agent system learns to control the components of the microgrid so as this to achieve its purposes and operate effectively, by means of a distributed, collaborative reinforcement learning method in continuous actions-states space. Stand-alone microgrids present challenges regarding guaranteeing electricity supply and increasing the reliability of the system under the uncertainties introduced by the renewable power sources and the stochastic demand of the consumers. In this article we consider a microgrid that consists of power production, power consumption and power storage units: the power production group includes a Photovoltaic source, a fuel cell and a diesel generator; the power consumption group includes an electrolyzer unit, a desalination plant and a variable electrical load that represent the power consumption of a building; the power storage group includes only the Battery bank. We conjecture that a distributed multi-agent system presents specific advantages to control the microgrid components which operate in a continuous states and actions space: For this purpose we propose the use of fuzzy Q-Learning methods for agents representing microgrid components to act as independent learners, while sharing state variables to coordinate their behavior. Experimental results highlight both the effectiveness of individual agents to control system components, as well as the effectiveness of the multi-agent system to guarantee electricity supply and increase the reliability of the microgrid.

[*] Corresponding author at: Department of Digital Systems, University of Piraeus, Piraeus, Greece.
E-mail address: panagiotis.kofinas@gmail.com (P. Kofinas).

https://doi.org/10.1016/j.apenergy.2018.03.017
Received 25 December 2017; Received in revised form 3 March 2018; Accepted 15 March 2018
Available online 21 March 2018

Nomenclature

MAS	multi-agent system
FLS	Fuzzy Logic System
D_m	fuzzy set of inputs
c	output variable of fuzzy rule
w_i	firing strength of rule i
$(\cap)$	intersection operator
a	global output/action
S_i	fuzzy sets of state variables
η	learning rate
R	reward
q	q-value of rule
AG	set of agents
T	state transition function
MDP	Markov Decision Process
f	weight function
pwt	percentage water in the tank
ed	water demand of electrolyzer (l/h)
P_L	demanded power of the variable electrical load (W)
R_{DA}	reward of desalination agent
SOC	state of charge
P_{BC}	battery charge power (W)
R_{BAT}	reward of battery agent
p_{H_2}	percentage of hydrogen in the tank
d_{H_2}	demanded hydrogen of fuel cell (m^3/h)
α_{bat}	control signal of the battery agent
P_{FC}	power produced by the fuel cell (W)
P_{DG}	power produced by the diesel generator (W)
DC	Direct Current
V	cumulative expected discounted reward
PV	photovoltaic
RL	Reinforcement Learning
$\boldsymbol{x}$	crisp input/state vector
E	output fuzzy set defined by the expert
$(\cup)$	union operator
a_i	consequent/action of rule i
TSK	Tagaki-Sugeno-Kang
X_i	set of state variables
γ	discount factor
FIS	fuzzy inference systems
t	set of discrete time points
A	set of discrete actions
p	transition probability
Q	Q-function
ANFIS	neuro fuzzy inference system
wd	water demand (l/h)
P_{PV}	photovoltaic potential power production (W)
pbdesalination	power balance for desalination agent (W)
P_{des}	power consumption of the desalination unit (W)
pb_Battery	power balance for battery agent
P_{BD}	battery discharge power (W)
L_p	percentage of the demanded power of the dynamic electrical load
pb_Electrolyzer	power balance for electrolyzer agent (W)
R_{EA}	reward of electrolyzer agent
R_{FCA}	reward of fuel cell agent
R_{DG}	reward of fuel cell agent
MF	membership function
Π	policy
E	expectation operator

1. Introduction

1.1. Microgrids and control

For several decades, the power production is based on a central system with large scale conventional power plants and extended power transmission networks with lack in flexibility and extensibility [1]. Nowadays, the trend in power generation is changing and shifting to the distributed power generation paradigm [2]. This new model allows incorporation of new technologies with low or zero emission of gasses which do not affect the environment [3].

Microgrids are usually low voltage networks with distributed power generation units, storage devices and controllable loads [4]. They have clearly defined electrical boundaries that act as single controllable entities with respect to the grid [5]. Microgrids can operate in either grid-connected or island-mode [6]. Their ability to operate in island-mode makes them an ideal solution in remote areas, rural areas and islands [7] where the grid expansion is either impossible or cost prohibitive [8].

The ability of operating in grid-connected mode makes them an efficient economic solution in power market [9]. Thus, microgrids are exceptional infrastructure for serving the current trend of distributed power generation [10–11]. On the other hand, despite the benefits provided by the microgrid architecture there are some challenging tasks. The most challenging task is the energy management of the microgrid. In grid connected mode, in many cases, the energy management has to deal with economic problems. The schedule of the energy storage and use has to be optimal, in order to maximize the economic benefits under the dynamic prices of the electricity market.

In island mode, the main challenge is to guarantee electricity supply and maintain (or increase) reliability of the microgrid under the uncertainties which are introduced by the renewable power sources and the stochastic demand of the consumers. This becomes even more challenging when the number of renewable power sources and dynamic loads increase [12]. A centralized management and control system presents limitations, requiring distributed sources and loads to communicate their state to the central controller, while the control actions have to be broadcasted back to each unit [13–14]. In doing so, given components' possible states, the number of global system states increase exponentially to the number of components, which is also the case for the combination of components' control actions [15]. Additionally, failure of the central controller decreases the reliability of the system. The aforementioned limitations can be addressed by applying a decentralized control method. The computational load is shared among the local controllers of each system components, while the reliability of the system increases, since a failure in the local controller may not affect the whole system's performance [16]. A considerable benefit of decentralized control is that new components may be added seamlessly to the whole system, or existing components may be replaced with new ones, given that their controllers satisfy information sharing requirements for the whole system to operate successfully.

1.2. Microgrid and multi-agent system (MAS)

A multi-agent system consists of a group of agents that interact with each other and with their environment [17]. This system is ideal for solving complex problems by factoring the problem to a number of smaller and simpler ones that can be solved in more computational efficient ways than using a single-agent system. Additionally, it provides solutions that respects the autonomy of components (e.g. each component has different operating preferences, constraints, etc.). These features make multi-agent systems ideal for solving energy management problems [18].

MAS have been previously used by researchers to deal with the task of energy management in buildings [19] and expanded to the management of microgrids [20–21]. Furthermore, MAS have been used in microgrids with distributed generators, loads and electrical vehicles in order to ensure coordinated power management through effective utilization of electrical vehicles [22–23]. Additionally, MAS have been proposed for optimal management of buying and selling power between microgrids and electrical grid aiming to economic viability and maximization of the economic benefit [24–26].

Many researches present MAS for energy management in grid connected and stand-alone microgrids. In grid connected microgrids the MAS focus on the smooth operation in transition between island mode and grid connection and on keeping stable the critical loads operation and therefore the whole microgrid operation. A MAS system to control power flow via the use of an auction algorithm is presented in [4]. The control scheme consists of three types of agents, the local agents which control the local units, the microgrid management agents that take decisions according to the state of the microgrid and the ancillary agents that only share information about the state of the electrical grid. These agents do not have adaptation or learning abilities and decide their action by the local information.

An intelligent distributed autonomous power system for controlling critical loads through a standardized interoperability algorithm is proposed in [27]. The MAS has four types of agents that have to collaborate in order to secure the critical loads operation. In [28] authors propose a MAS with a central coordinator for optimal response in emergency power demand. This work focuses on the hardware implementation of the agents using microcontrollers and the Zigbee wireless communication technology for communication among them. An intelligent coordinated control of a microgrid in both grid-connected and island modes using MAS, is proposed in [29]. The proposed method has local agents that aim to keep the frequency within the operating limit without shedding the loads and performing a smooth transition from island to grid tied operation.

In stand-alone microgrids most works on MAS focus on the power balance between the distributed generators and the controllable loads. In [20], a hybrid multi-agent based energy management system (composed by various types of agents) is proposed. This system has a central coordinator agent for distributed power generators and controllable loads. In the proposed scheme, the agents have to cooperate in order to reach coordination of energy and comfort management in integrated buildings and a microgrid system. An agent-based decentralized control model for island-mode microgrids is proposed in [15]. The control approach is a two-layer control structure. There is the bottom layer which is the microgrid and an upper layer which is the communication layer populated by the agents. Each agent collects the states of a distributed generator and the states of the load which is connected through the communication lines between the two layers, and then shares the information with its neighboring agents. The information is processed according to the predefined knowledge base and the agents adjust the output power of the generators at the next time step in order to balance the power. In addition to this, a multi-agent system for energy management of a Photovoltaic small hydro hybrid microgrid at high altitude in order to achieve efficient and stable system operation is proposed in [30]. The MAS has seven different types of agents with predefined capabilities that have to cooperate in order to achieve stable system operation. A virtual bidding strategy in combination with the dynamic power dispatch, secures the smooth operation of the microgrid. In [31] authors present a decentralized MAS that provides a cooperative (based on game theory) control framework for the power management in the autonomous microgrid systems. The system consists of different agents each one associated with a microgrid unit. In order to reduce the communication burden and avoid conflicts between agents, central facilitators are used. An autonomous MAS for power generation scheduling and demand side management has been proposed in [32–33] where the intelligent agents have their own knowledge base.

MAS systems have been used for balancing the generation and the demand in microgrids. Each agent communicates and shares information only with its neighboring agents reducing the computational burdens and complexities [22]. In [34] a MAS that is responsible for the management of renewable energy resources and power storage systems is presented, in order to minimize disturbances to the supply-and-demand balance. An algorithm decides which resources are connected and/or disconnected to the network and the MAS decides the power flow of each resource. For the communication between the agents a central agent is used for gathering the information. Each agent has discrete states and actions and the learning mechanism is based on distributed value function method. In [35], a decentralized control system is proposed in order to ensure the smooth operation and stability of the islanding microgrid. The control scheme is the distributed cascaded Proportional-Integral controller scheme for each distributed generator unit region. An advanced Lyapunov based technique is applied to the model, in order to prove stability and convergence to the desired equilibrium. In [36], a distributed control scheme for assuring stability in stand-alone microgrids with bus connected topology is proposed. There is a lower level of local controllers that stabilize the microgrid with a plug and play algorithm and a second layer with distributed communication controllers for further performance improvement. These two layers constitute a hierarchical control architecture.

Despite the variety of MAS approaches in the literature, some of the works aim to solve the trades-off between microgrids and the electrical grid regarding economic factors. Other works deploy a MAS structure for controlling the power balance both in island and/or grid connected modes with intelligent agents without an adaptation or any learning mechanism and some of them are just enhanced by off-line training algorithms. In these cases, changes to the topologies or the units of the microgrid can lead to power unbalances and failures to controlling the system successfully. Furthermore, the MAS approaches are not always fully decentralized and some coordinating agents are used. The use of central coordinator has the disadvantage that failure occurrence in the central controller leads to miscoordination or no coordination of the local controllers and chaotic behavior of the system in total. A few papers propose MAS with learning abilities, but most of them do not present a solution to deal with a continuous state-action space. Making a coarse-grained discretization in the continuous state-action space leads to loose of information, while performing fine-grained discretization in the continuous state-action space has as result the increase of the state space and state-action combinations.

This article proposes a collaborative multi-agent system for managing the energy in a stand-alone microgrid. Our objective is to increase the reliability of the system by balancing the power between the units of the microgrid and simultaneously cover the power demand, as well as the demand for goods and services, with the minimum use of fossil fuels (use of diesel generator). The MAS is fully distributed and the agents learn an optimal policy through reinforcement learning. Each agent is "equipped" with fuzzy Q–Learning in order to cope with the continuous state-action space. The microgrid considered consists of energy production units (Photovoltaic source, fuel cell and diesel generator), consumption units (electrolyzer, desalination plant and a dynamic energy load which represents the dynamic energy needs of the inhabitants) and a storage unit (Battery bank). Each agent is associated with one unit and provides control signal for adjusting the power flow. The agents are independent Q-learners, sharing information about their state according to their interactions, in order to function effectively the whole system, achieving its purposes. We use fuzzy Q-Learning techniques in order to cope with the continuous states and actions space. This is deemed necessary, as the variables that define the state of each unit are continuous and we must support continuous control signals in each unit.

The main contributions of this paper are as follows:

- We deploy a MAS with a modified Independent Learners approach to solve the complex problem of energy management in a stand-alone (island-mode) microgrid. The learning approach exploits local rewards and state information that are relevant to each agent. This has as result, the reduction of the states space and the enhancement of the learning mechanism.

- In order to cope with the continuous state and actions space, we introduce Fuzzy Q-Learning in each agent.
- The problem of power balancing between production and consumption units is addressed in a fully decentralized framework. Additionally, we address the problem incorporating consumer units that are responsible for providing services and/or goods at specific levels of demand. This provides more tradeoffs beyond pure parameters for energy production and consumption. For this purpose, as in previous work [8,37], we use a desalination unit.
- Experimental results are based on real data, concerning the demanded load and the energy produced by a photovoltaic source.

The structure of this paper is as follows: Section 2 provides preliminaries about Fuzzy Logic Systems. Section 3 provides preliminaries about Reinforcement Learning, Q-Learning, fuzzy Q-Learning and multi-agent systems. Section 4 presents the microgrid and the problem specification. Section 5 presents the design of the multi-agent system and justifies the choices made. Section 6 presents and discusses experimental results based on the simulated grid and finally, Section 7 concludes the paper sketching future work.

2. Elements of FLS

Fuzzy logic is the computational tool which allows complex processes to be expressed in general terms without the usage of complex models [38]. In classical set theory an element either belongs to a crisp set or not, while in fuzzy theory it might belong to a set, up to a degree. The extent in which an element participates in a fuzzy set is called the degree of membership and it is defined by a number in the range of $[0, 1]$. One of the strengths of fuzzy logic systems is the use of fuzzy if/then rules which express relations among fuzzy variables using linguistic terms [38]. Such rules are of the generic form:

Ru ***if*** (x_1 is D_m) and/or (x_2 is D_m)....and/or (x_m is D_m) ***then*** (c is E),*where* D_m is a fuzzy set of inputs, $\mathbf{x} = (x_1,x_2,...,x_m)$ is the crisp input vector, c is the output variable and E is a fuzzy set defined by the expert. The operators and/or combine the conditions of the variable inputs that should be satisfied arising the firing strength $w_i(\mathbf{x})$ of the rule. The operator and (respectively or) is represented with intersection (∩) (respectively, union (∪)). The standard intersection is defined with the "min" operator and alternatively with product (∗), while the standard union is defined with "max" operator and alternatively by the "probabilistic or" operator [39]. The firing strength $w_i(\mathbf{x})$ of the rule i is applied to the fuzzy set in the consequence of the rule resulting to a reshaped fuzzy set (fuzzy implication). The most common fuzzy implications are implemented by the Mamdani and Larsen implications via the "min" and the product (∗) operators respectively. The fuzzy sets that are the outcomes of rules are aggregated to form a single fuzzy set. The aggregation method can be implemented by the "max" operator or by the "probabilistic or" operator. A defuzzification method produces a crisp output value from the fuzzy sets after the aggregation. There are a lot of defuzzification methods with the most common to be the "centroid" calculation and the "weighted average" [40].

The aforementioned procedure can be specified using four blocks which constitute the block diagram of a FLS (Fig. 1). The first block is the fuzzifier which converts the crisp values of the input vector into fuzzy values. The second block is the knowledge base & data base which defines the membership functions and stores the fuzzy rules. The third block is the fuzzy inference engine, which performs the approximate reasoning exploiting the fuzzy rules, and finally, the weighted average/defuzzifier block that calculates the crisp output, given the fired rules.

The global output of a fuzzy system can be calculated by the Wang-Mendel model [40]. In this model the "and" operator and the fuzzy implication are implemented by the product operator and the defuzzification method is done by means of the weighted average [40]:

$$a(x) = \frac{\sum_{i=1}^{N} w_i(x)a_i}{\sum_{i=1}^{N} w_i(x)} \tag{1}$$

where a_i is the consequent of rule i which can be a fuzzy singleton in the simplified model of Tagaki-Sugeno-Kang or the centroid of a fuzzy set [41].

3. Reinforcement learning

The objective of reinforcement learning is to find a policy, i.e., a mapping from states to actions that maximizes the expected discounted reward [42], by learning through exploration/exploitation in the space of possible state-action pairs. Actions that share good performance when performed in a given state are "rewarded" and actions leading to a poor performance are "punished", providing feedback to the system in order to learn the "value" of actions in different states.

3.1. Q-Learning

The Q-Learning [43] is a reinforcement learning method where the agent computes the Q-Function that estimates the future discounted rewards of actions applied in states. The output of the Q-Function for a state $\mathbf{x}$ and an action a is represented as $Q(\mathbf{x},a)$. In Q-Learning the Q value of each action $\boldsymbol{a}$ when performed in a state $\mathbf{x}$ can be calculated as follows:

$$Q'(\mathbf{x},a) = Q(\mathbf{x},a) + \eta(R(\mathbf{x},a,\mathbf{x}') + \gamma_{a}^{max}Q(\mathbf{x}',a) - Q(\mathbf{x},a)) \tag{2}$$

where $Q'(\mathbf{x},a)$ is the updated value of the state-action combination when the agent receives the reward after performing the action a in the state $\mathbf{x}$. Q learning assumes that the agent continues for state $\mathbf{x}$ by performing the optimal policy, thus ${}^{max}_{a}Q(\mathbf{x}',a)$ is the maximum value of the best action performed in the state $\mathbf{x}'$ (the state resulting after performing a in state $\mathbf{x}$). The learning rate η determines in which degree the new information overrides the old one [44] and the discount factor γ determines the importance of the future rewards [45]. Overall, the Q-

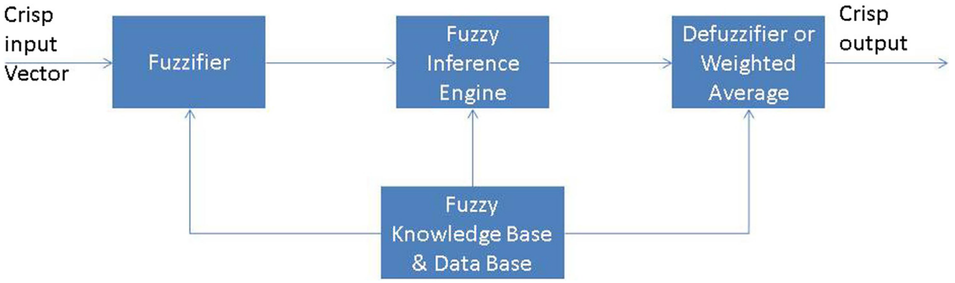

Crisp input Vector
Fuzzifier
Fuzzy Inference Engine
Defuzzifier or Weighted Average
Crisp output
Fuzzy Knowledge Base & Data Base

Fig. 1. Block diagram of an FLS.

Learning agent chooses the action a to be applied in the current state $\boldsymbol{x}$ and identifies the next state $\boldsymbol{x}'$. It gets the reward from this transition i.e. , and updates the value for the pair $(\boldsymbol{x},a)$ assuming that it performs the optimal policy from state $\boldsymbol{x}$ and on. A Q-Learning agent applies an exploration/exploitation scheme in order to compute a policy that maximizes payoff.

The main advantages of Q-Learning are that it does not need a model of the environment (in our case, the microgrid system), and can adequately handle problems involving stochastic transitions and rewards without requiring any specific adaptation. These characteristics, make Q-Learning suitable to be applied in our study, where a rigorous model of any microgrid is difficult to be built and adaptations are required according to needs and technology evolution. On the contrary, the Q-Learning can be inefficient for large states-actions spaces and cannot be applied in a straightforward manner to continuous action-state spaces that our problem involves. The simplest solution in order Q-Learning to be applied in continuous action-state spaces is to discretize the space. Discretizing the action-state space, may lead to sudden changes in actions even for smooth changes in states. Making the discretization more fine-grained will smooth these changes but the state-action combination increases. In problems with multidimensional state space the number of the combinations grows exponentially. In order to overcome the aforementioned limitations, fuzzy function approximation can be used in combination with Q-Learning. By this way, actions change smoothly in response to smooth changes in states without the need of discretizing the space in a fine-grained way [46].

3.2. Fuzzy Q-Learning

In Q-Learning, Q-values are stored and updated for each state-action pair, that if the number of state-action pairs is very large, the implementation might be impractical. Fuzzy Inference Systems have the advantage of achieving good approximations [47] in the Q-function and simultaneously make possible the use of the Q-Learning in continuous states-space problems (Fuzzy Q-Learning) [48]. In fuzzy Q-Learning, $\boldsymbol{x}$ is the crisp set of the inputs defining the state of the agent. These are converted into fuzzy values and each fuzzy rule corresponds to a state. In other words, the firing strength of each rule defines the degree to which the agent is in a state. Furthermore, the rules do not have fixed consequents; meaning there are no fixed (predefined) state-action pairs but through the exploration/exploitation algorithm arise the consequents of each rule (state-action pair). Thus, the FIS has competing actions for each rule and the rules have the form:

$$\begin{array}{lll} \textbf{\textit{if}}\ x\ \textit{is}\ S_i\ \textbf{\textit{then}} & & \alpha[i,1]\ \textit{with}\ q[i,1] \\ & \text{or} & \alpha[i,j]\ \textit{with}\ q[i,j] \\ & \vdots & \\ & \text{or} & \alpha[i,k]\ \textit{with}\ q[i,k] \end{array}$$

where $\alpha[i,k]$ is the k^{th} possible action in rule i and $q[i,k]$ its corresponding q value. The state S_i is defined by $(x_1\ \textit{is}\ S_{i,1}\ \textit{and}\ x_2\ \textit{is}\ S_{i,2}\ \ldots\textit{and}\ x_n\ \textit{is}\ S_{i,n})$, where $S_{i,j}$, $j=1,\ldots,n$ and $j=1,\ldots,n$ are fuzzy sets. Specifically, the algorithm of the fuzzy Q-Learning is presented below.

Fuzzy Q-Learning algorithm

```
Observe state $x$
Take an action $\alpha_i$ for each rule $i$ according to exploration/exploitation algorithm
Calculate the global output $a(x)$ according to Eq. (3)
Calculate the corresponding value $Q(x,a)$ according to Eq. (4)
Capture the new state information
Calculate reward
Update q-values according to Eq. (5)
```

(1) Observation of the state $\boldsymbol{x}$.
(2) For each fired rule one action is selected according to the exploration/exploitation strategy.
(3) Calculation of the global output $a(\boldsymbol{x})$ and calculation of the corresponding value $Q(\boldsymbol{x},a)$.

$$a(\boldsymbol{x}) = \frac{\sum_{i=1}^{N} w_i(\boldsymbol{x}) a_i}{\sum_{i=1}^{N} w_i(\boldsymbol{x})} \tag{3}$$

$$Q(\boldsymbol{x},a) = \frac{\sum_{i=1}^{N} w_i(\boldsymbol{x}) a_i q[i,i^\dagger]}{\sum_{i=1}^{N} w_i(\boldsymbol{x})} \tag{4}$$

where a_i is as specified in Section 2 (consequents of the rule i) and corresponds to the selected action of rule i, $q[i,i^\dagger]$ is the corresponding q-value of the fired rule i for the selection of the action $i^\dagger$.

(4) Application of the action $a(\boldsymbol{x})$ and observation of the new state $\boldsymbol{x}'$.
(5) Calculation of the reward .
(6) Updating the q values according to:

$$\Delta q[i,i^\dagger] = \eta \Delta Q \frac{w_i(\boldsymbol{x})}{\sum_{i=1}^{N} w_i(\boldsymbol{x})} \tag{5}$$

where $\Delta Q = R(\boldsymbol{x},a,\boldsymbol{x}') + \gamma V(\boldsymbol{x}',a^*) - Q(\boldsymbol{x},a)$, $V(\boldsymbol{x}',a^*) = \frac{\sum_{i=1}^{N} w_i(\boldsymbol{x}') a_i q[i,i^*]}{\sum_{i=1}^{N} w_i(\boldsymbol{x}')}$ and $q[i,i^*]$ is the selection of the action i^* that has the maximum Q value for the fired rule i.

3.3. MAS and Q-Learning

In many distributed problems the agents have to work in coordination in order to optimize a shared performance measurement [16]. Such a problem may be formulated as an extension of a Markov Decision Process (MDP) for an agent [49] comprising the following constituents:

- A set of discrete time points $t = t_0,t_1,t_2,t_3,\ldots$
- A set of agents $AG = \{AG_1,AG_2,\ldots,AG_n\}$.
- A set of state variables X_i. The global state is defined by the cross product of all m variable sets: $X = X_1 \times X_2 \times \ldots \times X_m$. A state $\boldsymbol{x^t} \in X$ that describes the state of the "world" at the time step t.
- A set of discrete actions A_i for any given agent i. The action that the agent i selects at the time step t is defined as $\alpha_i^t \in A_i$. The joint action $\boldsymbol{a}^t \in A = A_1 \times A_2 \times \ldots \times A_n$ is the combination of all individual actions of the n agents.
- A state transition function $T{:}\ X \times A \times X \to [0,1]$ that gives the transition probability $p(\boldsymbol{x}^{t+1}|\boldsymbol{x}^t,\boldsymbol{a}^t)$ of the environment to move from the state $\boldsymbol{x}^t$ to the state $\boldsymbol{x}^{t+1}$ when the joint action $\boldsymbol{a}^t$ is applied to the state $\boldsymbol{x}^t$.
- A reward function $R_i{:}\ X \times A \to \boldsymbol{R}$ which provides to agent i an individual reward $r_i^t \in R_i(\boldsymbol{x}^t,\boldsymbol{a}^t)$ based on the joint action $\boldsymbol{a^t}$ which is applied at the state $\boldsymbol{x^t}$.

Solving the MDP has as result the computation of a policy. The term policy means a decision making function $\pi{:}\ X \to A$ that defines which action the agent should take at any state. The target is to find an optimal policy π^* that maximizes the objective function of the cumulative expected discounted reward for each state $\boldsymbol{x}$:

$$V^*(\boldsymbol{x}) = max_\pi E\left[\sum_{t=0}^{\infty} \gamma^t R(\boldsymbol{x}^t,\pi(\boldsymbol{x}^t))|\pi,\boldsymbol{x}^0 = \boldsymbol{x}\right], \quad \boldsymbol{x} \in X \tag{6}$$

where the expectation operator $E[\cdot]$ averages over stochastic transitions. Given a multi-agent MDP formulation, and in order to solve any such problem, we can apply collaborative reinforcement learning methods [18]. Considering Q-Learning methods, there are four main approaches: The first one is MDP Learning where the MAS can be considered as an agent with multiple states and an action vector. The agent can learn the Q values by applying the single agent Q-Learning algorithm [18].

The second approach is the Coordinated Reinforcement Learning. In many cases, the problem may be factored by defining a coordination graph between agents [50]. Two agents are related in such a graph if they share common state variables and thus, the action of one affects the state of the other. Related agents are called “neighbors”. In such cases, each agent has to coordinate its action with its neighbor agents. The global Q function is analyzed as a linear combination of the Q_i functions of neighbor agents [51].

$$Q_i(x_i,a_i):=Q_i(x_i,a_i)+\eta[R(\boldsymbol{x},\boldsymbol{a},\boldsymbol{x}')+\gamma_{a'}{}^{max}Q(\boldsymbol{x}',\boldsymbol{a}')-Q(\boldsymbol{x},\boldsymbol{a})] \tag{7}$$

A third approach is the Distributed Value Function where each agent keeps a local Q function based on its own actions and states. This function may be updated by embedding the Q functions of the neighbor agents. In order to define which agent are neighbors, a graph structure of agent dependencies is established beforehand. A weight function is used $f(i,j)$ which defines how much the Q value of an agent j contributes to the updating of the Q value of the agent i [52].

$$Q_i(x_i,a_i):=(1-\eta)Q_i(\boldsymbol{x},a_i)+\eta[R_i(\boldsymbol{x},\boldsymbol{a},\boldsymbol{x}')+\gamma\sum_{j\in\{i\cup\Gamma(i)\}}f(i,j)_{a_j'}{}^{max}Q_j(\boldsymbol{x}',a_j')] \tag{8}$$

According to the Independent Learners approach each agent acts autonomously. In the independent learners approach agents act and learn their policies independently from the others: i.e. one does not know the Q values of others, neither their policies. Each agent stores and updates its own Q table. The global Q function is defined as a linear combination of all local contributions [53].

$$Q_i(\boldsymbol{x},a_i):=Q_i(\boldsymbol{x},a_i)+\eta[R_i(\boldsymbol{x},\boldsymbol{a},\boldsymbol{x}')+\gamma_{a_i'}{}^{max}Q_i(\boldsymbol{x}',a_i')-Q_i(\boldsymbol{x}',a_i')] \tag{9}$$

In Independent Learners approach, the environment is no longer stationary [54] and convergence cannot be guaranteed, as the agents view only their state and ignore the actions of the other agents. Despite this disadvantage, this approach has been successfully applied to many control problems [18]. Additionally, many modified approaches of this method using local rewards and/or state information that are relevant to each agent can be used for reducing the state space and enhancing the learning mechanism [55].

4. Microgrid description

Fig. 2 presents the stand-alone solar microgrid considered in our study. There are three types of units in the microgrid: power production units, power consumption units and power storage units, with the later operating as power production units or power consumption units. The power production group of units includes the PV source, the fuel cell and the diesel generator (in green color). The power consumption group includes the electrolyzer unit, the desalination plant and the variable electrical load (in red color). Only the battery bank is included in the power storage group. The main control problem in stand-alone microgrids is to keep a reliable power supply to consumers under the uncertainties of power production and power consumption.

4.1. Microgrid units description

The photovoltaic source is a 20 kW solar park associated by a maximum power point tracker in order to deliver the maximum available power. The diesel generator has nominal power of 2 kW and the fuel cell has nominal power of 3 kW. The consumption rate of H_2 of the fuel cell equals to 39 l/min, operating at nominal power. The variable electrical load in our case represents a small residential building of four households with peak load of 10 kW. The desalination unit has potable water production rate of 106 l/h operating at its maximum power of 613 W and the electrolyzer unit has nominal power of 7.4 kW. The water consumption rate of the electrolyzer unit equals to 0.8 l/h and the H_2 production rate equals to 1 m^3/h. The battery bank capacity equals to 30 kWh with maximum charge and discharge power of 10 kW for protection reasons. Deep discharges of the battery as well as high charge and discharge rates have as result the reduction of the battery lifetime [56]. In the microgrid, a water tank of 1 m^3 for water storage is additionally adopted and H_2 tank of 18 m^3 (compressed cylinders) for hydrogen storage. The model of the desalination plant is a data driven model developed by applying the adaptive Neuro Fuzzy Inference System (ANFIS) technique in our previous work [57]. The other models which are used are linear models. This is for simplicity reasons as the way that the model changes its output in relation to the input does not

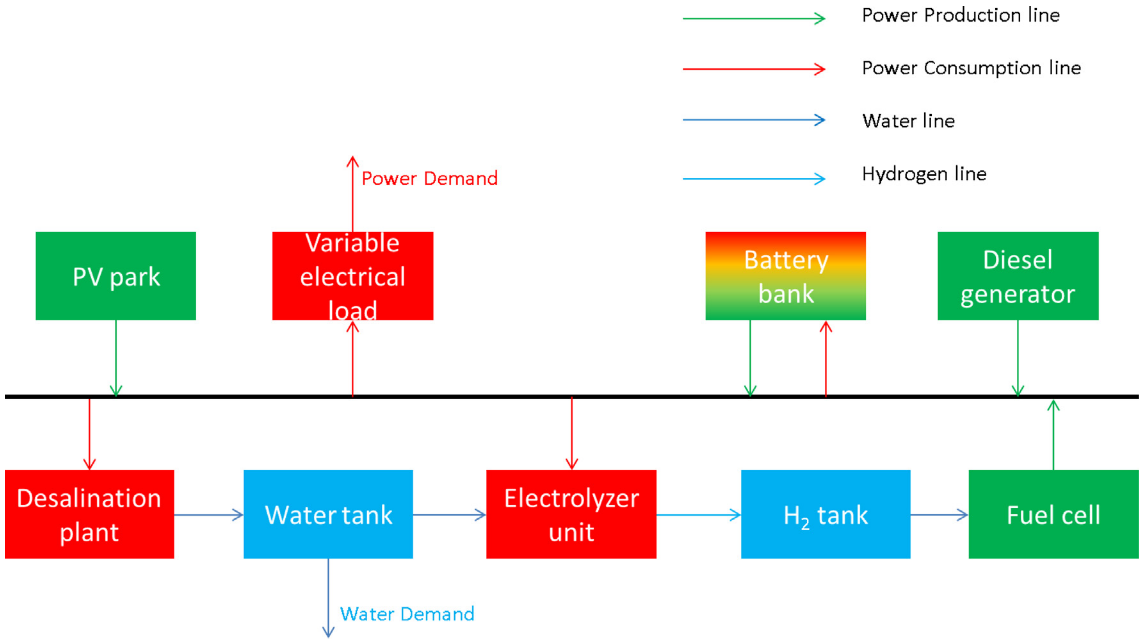


Fig. 2. Overview of the solar microgrid (green and red colors indicate the power production and power consumption units respectively, while the blue color indicates the water and hydrogen tanks). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

affect the energy management strategy.

4.2. Microgrid analysis and constraints

The main power supply of the microgrid is the PV source. Thus, the exploitation of the PV generated power is a priority. If the power of the PV source exceeds the total power demand, then the total demand is covered exclusively by the PV source. If the total demanded power exceeds the total generated power, then the produced power is delivered to the consumption units proportionally to their demanded power. Each unit is associated with one agent that controls either the amount of power delivered to the microgrid for production units, or the amount of power consumed for consumption units, or both, the amount of power delivered and consumed for storage units.

Neither the PV source nor the dynamic load have an associated agent. It must be pointed out that the PV source, as an inexhaustible source, has to produce the maximum feasible power and no agent is associated with it. The dynamic load represents the inhabitants' power needs and the objective is to supply all the amount of power that is demanded. For the rest of the units, each associated agent produces a control signal representing the percentage of the power to be produced or consumed according to their nominal or demanded power. The structure of the MAS, the interactions among agents, and each individual agent are presented extensively in the next chapter.

5. Fuzzy MDP framework and MAS

5.1. Fuzzy MDP framework

According to the architecture of a cooperative MAS, the system can be presented as an extension of an MDP for one agent. In our case this MDP is composed by:

- A set of discrete time points $t = t_0,t_1,t_2,t_3,...$ with $t_{i+1}-t_i = 50$ s.
- A group of five agents $A = \{AG_1,AG_2\ AG_3\ AG_4\ AG_5\}$ where the AG_1 is the desalination agent, AG_2 is the battery agent, AG_3 is the electrolyzer agent, AG_4 is the fuel cell agent and AG_5 is the diesel generator agent.
- A group of fuzzy state variables X_i. The global state is defined by the cross product of all thirteen variables: $X = X_1 \times X_2 \times ... \times X_{13}$. Three variables for each of the desalination, the electrolyzer and the fuel cell agent, and two variables for each of the battery and the diesel generator agent. A state $\boldsymbol{x}^t \in X$ that describes the state of the "world" at the time step t.
- A group of fuzzy actions (fuzzy singletons) A_i for any given agent i. For the AG_1 (desalination) agent there is a group of six fuzzy singletons $A_1 = \left\{\frac{1}{0} + \frac{1}{0.2} + \frac{1}{0.4} + \frac{1}{0.6} + \frac{1}{0.8} + \frac{1}{1}\right\}$. Where "+" is the union operator and "−" denotes a particular membership function to a value on the universe of discourse. For the AG_2 (battery) agent there is a group of seven fuzzy singletons $A_2 = \left\{\frac{1}{-0.5} + \frac{1}{-0.1} + \frac{1}{-0.01} + \frac{1}{0} + \frac{1}{0.01} + \frac{1}{0.1} + \frac{1}{0.5}\right\}$ and for the AG_3 (electrolyzer), AG_4 (fuel cell) and AG_5 (diesel generator) agents there is a group of five singletons for each one $A_3 = \left\{\frac{1}{-0.5} + \frac{1}{-0.1} + \frac{1}{0} + \frac{1}{0.1} + \frac{1}{0.5}\right\}$, $A_4 = \left\{\frac{1}{-0.1} + \frac{1}{-0.01} + \frac{1}{0} + \frac{1}{0.01} + \frac{1}{0.1}\right\}$ and $A_5 = \left\{\frac{1}{-0.1} + \frac{1}{-0.01} + \frac{1}{0} + \frac{1}{0.01} + \frac{1}{0.1}\right\}$. The joint action $\boldsymbol{a}^t$ is the combination of all individual actions of the five agents at any specific instant t.
- The control problem is a deterministic problem since the transitions will be the same (T: $S \times A \rightarrow S$) for every state-action combination under the same conditions of production and consumption.
- A reward function R_i: $X_i \times A_i \rightarrow \boldsymbol{R}$ which provides to agent i an individual reward $r_i^t \in R_i(\boldsymbol{x}^t,\boldsymbol{a}^t)$ based on the joint action $\boldsymbol{a^t}$ which is applied at the state $\boldsymbol{x^t}$. The reward functions of the agents are described in the next sub-chapter.

5.2. MAS

Fig. 3 shows the MAS overview. The blue arrows represent the input signals for defining the state of the agent while the red arrows represent the control signals produced by the agents in order to control the associate unit. Blue arrows and corresponding agents define the

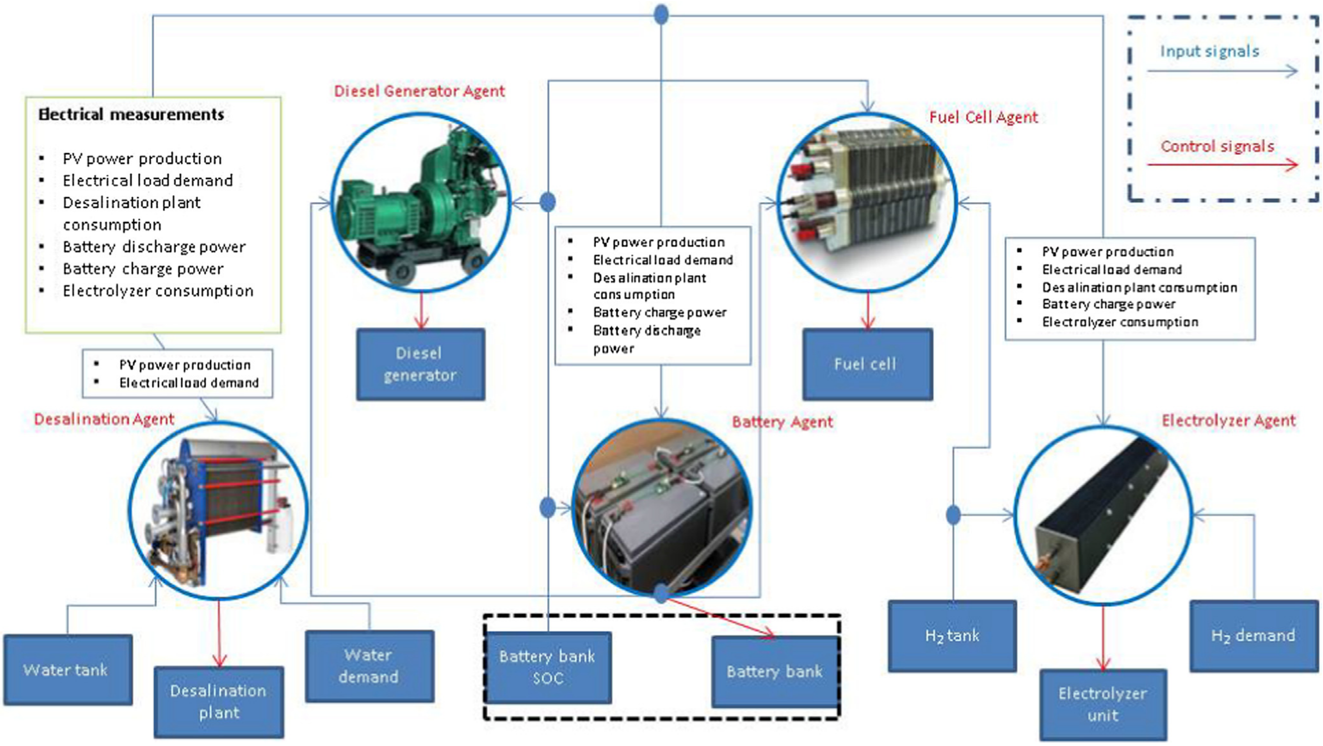


Fig. 3. MAS overview.

coordination graph for our multi-agent system. The input signals of the agents define the values of their state variables and are conditioning in the range of [0, 1] for signals with positive values and in the range of [−1, 1] for signals with both positive and negative values.

For each input with positive values we use five membership functions (MFs) and for each input with both positive and negative values we use seven MFs (Fig. 4). The PVB, PB, PM, PS, Z, NS, NM and NB denote Positive Very Big, Positive Big, Positive Medium, Positive Small, Zero, Negative Small, Negative Medium and Negative Big, respectively.

The agents take into account only those quantities arising by electrical measurements in the microgrid which are relevant for them in order to act properly and in hierarchical manner (the input variables and the states of the agents are described extensively below). This hierarchical manner guarantees that agents act in a fully coordinated way w.r.t. priorities to the units operation. In this case, regarding the consumption units, the water demand precedes over the electrical load, the electrical load precedes over the battery charging and the battery charging precedes over the hydrogen storage. The priority list concerning the production units has as follows: first the battery in discharge mode, second the fuel cell and last the diesel generator. Having said that, it must be pointed out that we do not impose any hard-wired rules or structure to the system, nor any specific coordinator agent, but we "impose" that behavior by shaping the reward functions of agents.

The desalination agent only needs to know if the power produced by the PV source is greater than the power consumed by the load. By this way is given a virtual priority first to the dynamic electrical load and then to the desalination unit.

The battery agent needs to know the power balance between the PV source, the electrical load, the desalination and its own power production/consumption. By this way, the battery serves both the electrical load and the desalination unit and simultaneously checks the power flow of the battery for avoiding power deficit due to the battery charging power.

The electrolyzer agent needs to know the power balance between the PV source, the electrical load, the desalination, the battery power consumption and its own power consumption. Thus, the electrolyzer unit grants the priority regarding the power consumption to the electrical load, to the desalination unit and to the battery. The battery power production is not taken into account, in order to avoid confusions regarding the power surplus. For example, in the case where the battery power production is taken into account, the electrolyzer agent may detect surplus to power due to the battery discharge. By this way, the electrolyzer unit will consume power from the battery in order to produce Hydrogen. Subsequently, the hydrogen will be used by the fuel cell to produce power. Thus, saved energy is consumed in order to produce a smaller amount of energy due to thermal losses between the energy conversions.

The fuel cell agent and the diesel generator agent have as input the control signal of the battery. This input defines how much and when the battery is charging or discharging. These two units act ancillary to the battery and have to coordinate their actions according to the battery state to avoid situations that lead to energy waste. For example, if the SOC of the battery is low and the battery is charging there is no need for these units to operate.

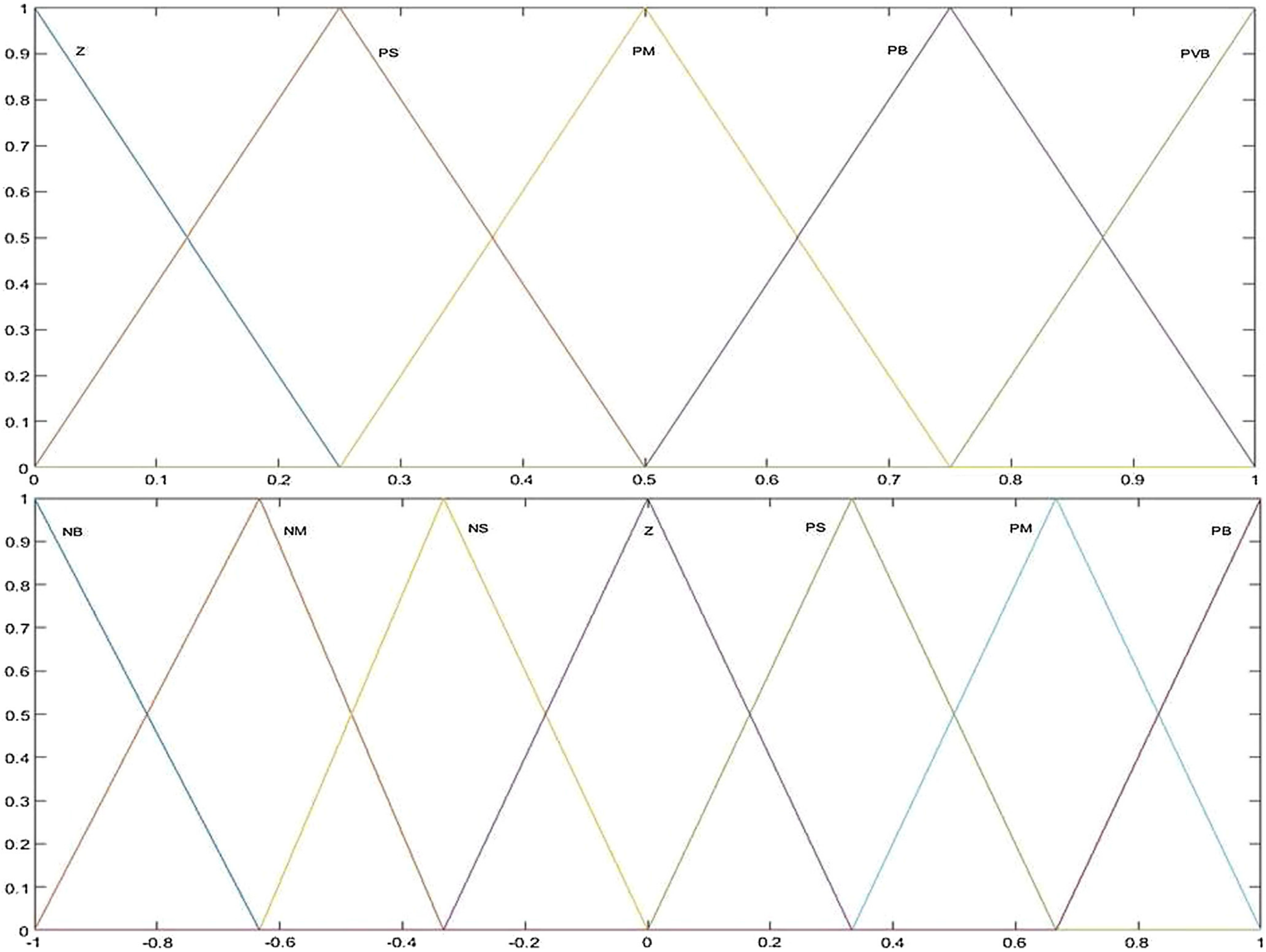


Fig. 4. Input membership functions.

All the agents are using the same exploration/exploitation scheme. In order to perform high exploration when the agent goes to a new state, agents explore for a certain number of rounds per state (500 rounds/state in our case study) and then check and performs the action/s that has/have not been performed at all (if there are any). This allows all possible actions to be explored for each state. After that, the agent performs exploitation for 99% and 1% exploration for a given state. The global output of each agent is calculated by the Wang-Mendel model and the rule base consists of TSK rules that have fuzzy singleton sets in the consequents of the rules. The fuzzy singleton sets of the output vector provided by each agent represent different levels of the continuous control signal. The final control signal arises by applying the Weighted Average defuzzification method.

5.2.1. Desalination agent

The desalination agent has three variable inputs as follows:

- The percentage water in the tank (*pwt*) which is in the range [0, 1];
- the water demand (*wd*) which equals to the sum of the habitants water demand and the electrolyzer demand (*ed*), both normalized in the range [0,1]; and
- the power balance (*pb_desalination*) between the photovoltaic potential power production (P_{PV}) and the demanded power of the variable electrical load (P_L), i.e. $pbdesalination = P_{PV} + P_L$, normalized in the range [−1, 1].

There are two inputs with five MF's each, and one input with seven MFs, which result to 175 states represented by an equal number of rules. The output vector of the desalination agent has six fuzzy singleton sets and the global action defines the percentage of the power to be consumed by the desalination unit according to its nominal operating power.

The desalination agent reward R_{DA} is defined as follows:

$$R_{DA}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{x}') = \begin{cases} pwt(\boldsymbol{x}')-pwt(\boldsymbol{x}) & pbdesalination \geqslant 0 \\ 0.5-(P_{des}(\boldsymbol{x}'))/613 & pbdesalination < 0 \ \& \ (pwt(\boldsymbol{x}') \\ & -pwt(\boldsymbol{x}')) \geqslant 0 \\ -(P_{des}(\boldsymbol{x}'))/613 & pbdesalination < 0 \ \& \ (pwt(\boldsymbol{x}') \\ & -pwt(\boldsymbol{x}')) < 0 \end{cases} \tag{10}$$

where P_{des} is the power consumption of the desalination unit.

In the first case, when there is surplus in the power balance, the reward depends on the change of the volume of water into the tank. In the second case, when there is a deficit in power balance, the reward depends on keeping the change of the water volume positive and simultaneously minimizing the power consumption, aiming to find the optimum operation point between the power consumption and the change of the water volume. In the third case the rewards depend only on power consumption aiming to find an operation point that leads the agent out of this case. In second and third case the division of P_{des} with the number "613" (which is the nominal power of the unit) is made for normalizing the reward in the range of [−1, 1] while in the first case the reward is in these limits.

5.2.2. Battery agent

The battery agent has two variable inputs which are:

- The SOC of the battery which is in the range [0, 1]; and
- the power balance (pb_Battery) between the P_{PV}, the P_L, the desalination consumption (P_{DES}) and the battery charge/discharge (P_{BC}/P_{BD}) power normalized in the range [−1, 1]

$$pbBattery = P_{PV} + P_L + P_{DES} + P_{BC} + P_{BD} \tag{11}$$

There is one input with five MF's and one input with seven MFs, this results to 35 states represented by an equal number of rules. The output vector of the battery agent has seven fuzzy singleton sets and the global action defines the percentage of the power change to either be consumed or produced by the battery bank according to its maximum charge and discharge power. The Battery Agent Reward (R_{BAT}) is defined as follows:

$$R_{BAT}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{x}') = (L_p + (SOC(\boldsymbol{x}')-SOC(\boldsymbol{x})))/1.2 \tag{12}$$

where L_p indicates the percentage of the demanded power of the dynamic electrical load that has been covered. When the battery is in discharge mode the battery agent should provide to the dynamic electrical load its demanded power in order the L_p to take its maximum value of "1". In case the change of the battery SOC is negative, to maximize the R_{BAT}, the discharging power of the battery has to be as much lower as to cover the load demand. If the battery is in charge mode then it tries to raise the SOC of the battery without making a deficit in the delivered load power which will lead to reduction of the R_{BAT} due to the reduction of the L_p.

5.2.3. Electrolyzer agent

The electrolyzer agent has three variable inputs which are:

- The percentage of H_2 in the tank (p_{H_2}) which is in the range [0, 1];
- the H_2 demand (d_{H_2}) by the Fuel cell normalized in the range [0, 1]; and
- the power balance (pb_Electrolyzer) between the P_{PV}, the P_L, the P_{DES}, the P_{BC} and the electrolyzer power consumption P_E normalized in the range [−1, 1]

$$pbElectrolyzer = P_{PV} + P_L + P_{DES} + P_{BC} + P_E \tag{13}$$

There are two inputs with five MF's each and one input with seven MFs, which results to 175 states represented by an equal number of rules. The output vector of the electrolyzer agent has five fuzzy singleton sets and the global action defines the percentage of the power change that the electrolyzer unit consumes according to its nominal operating power (7400). The Electrolyzer Agent Reward R_{EA} is defined as follows:

$$R_{EA}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{x}') = \begin{cases} 3*(p_{H_2}(\boldsymbol{x}')-p_{H_2}(\boldsymbol{x})) & ,pbElectrolyzer\ gt; 0 \\ -P_E/7400 & ,pbElectrolyzer \leqslant 0 \end{cases} \tag{14}$$

In the first case where the pb_Electrolyzer is positive, the R_{EA} of the electrolyzer agent is rising when the volume of the H_2 is rising. The multiplication with the "3" is made for conditioning the reward in the range of [−1, 1]. On the contrary, in the second case where the pb_Electrolyzer is zero or negative the operation of the electrolyzer unit will lead to even more deficit in power balance. The division with the nominal power of the unit is made for conditioning the reward in the range of [−1, 1].

5.2.4. Fuel Cell Agent

The Fuel Cell Agent has three variable inputs which are:

- The SOC of the Battery which is in the range [0, 1];
- the p_{H_2} which is in the range [0, 1]; and
- the control signal of the battery agent (α_{bat}) which is in the range of [−1, 1].

There are two inputs with five MF's each and one input with seven, this results to 175 states represented by an equal number of rules. The output vector of the Fuel cell agent has five fuzzy singleton sets and the global action defines the percentage of the power change to the produced power by the fuel cell according to its nominal power. The Fuel Cell Agent Reward (R_{FC}) is defined as follows:

$$R_{FCA}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{x}') = \begin{cases} P_{FC}/3000 & ,\alpha_{bat} < 0 \text{ and } SOC < 0.7 \\ -P_{FC}/3000 & \text{else} \end{cases} \quad (15)$$

where P_{FC} is the power produced by the fuel cell. In the first case where the SOC of the battery drops under 70% and the battery is discharging, the R_{FCA} is rising as the delivered power from the fuel cell is rising. This has as result the supply of extra power to the microgrid in order to rise the battery discharge time. In the second case, operating the fuel cell leads to negative values of R_{FCA} as there is no need of operating the unit either because the battery is charging or because the SOC is in high levels and the battery does not being in risk of a deep discharge. In both cases the division with the number "3000" (nominal power of the unit) is made for conditioning the reward in the range of $[-1, 1]$.

5.2.5. Diesel generator agent

The diesel generator agent has two variable inputs which are:

- The SOC of the battery which is in the range $[0, 1]$,
- and the α_{bat} which is in the range of $[-1, 1]$

These two inputs have five MF's each and the total amount of rules equals to 35 states represented by an equal number of rules. The output vector of the diesel generator agent has five fuzzy singleton sets and the global action defines the percentage of the produced power change by the diesel generator according to its nominal power. The diesel generator agent Reward (R_{DG}) is defined as follows:

$$R_{DG}(\boldsymbol{x},\boldsymbol{a},\boldsymbol{x}') = \begin{cases} P_{DG}/2000 & ,\alpha_{bat} < 0 \text{ and } SOC < 0.4 \\ -P_{DG}/2000 & ,\text{else} \end{cases} \quad (16)$$

where P_{DG} is the power produced by the diesel generator. The diesel generator agent tries not to fall the battery SOC in low level by providing the appropriate amount of power when the battery is discharging. In any other case, the operation of diesel generator is prohibited as for providing power uses fossil fuel. The division with the number "2000" (nominal power of the unit) is made for conditioning the reward in the range of $[-1, 1]$.

6. Simulation results

The simulation time is set to one year with a simulation step time of 5 s. The data, concerning the power production, is acquired from a 20 kW PV park located in Attica Greece with sample time of 300 s. The power consumption data set is collected from a four-household building with 50 s sampling time. For the water demand, due to lack of real data, a constant consumption of 120 l/h is assumed, with a variation of 40 l/h for 12 h/day (during daytime) and 0 l/h with a positive variation of 20 l/h for 12 h/day (during night). During the day, the variation follows a uniform distribution ranging from $[-40, 40]$ lt/h while during the night the variation follows a uniform distribution ranging from $[0, 40]$ l/h [37]. Fig. 5 presents the power production of the PV source per week and Fig. 6 presents the demanded power of the load per week.

According to the results provided, the whole performance of the MAS is as follows:

- The power balance between production and consumption is almost stabilized to zero (Fig. 7a). This results to low level of energy that has not been covered. The total energy consumption and the individual energy consumption of the units can be seen in Fig. 7b. The total amount of uncovered energy is very low and equals to 327.11 kWh, this quantity corresponds to only 1.54% of the total demand. Considering that the system seems to converge to a policy after the first 3.5 months (Fig. 7) the amount of the uncovered energy after the convergence becomes even lower. The amount of the uncovered energy until the convergence is 304.47 and equals to 93.1% of the total uncovered energy while the total uncovered energy from the convergence until the end of the year is only 22.64 kWh which corresponds to 6.9% of the total uncovered energy. Details for both the total and the individual energy consumptions are provided in Table 1.
- The uncovered demand of the water and the Hydrogen is very low (Fig. 8). The uncovered demand of the water is 350.1 l and located in the beginning of the year where the responsible agent performs mostly exploration. The same scenario is followed by the uncovered demand of the Hydrogen where the uncovered demand equals to 1.891 m^3 (Fig. 8) and located during the exploration phase. Additionally, the SOC of the battery drops under 20% (deep discharge)

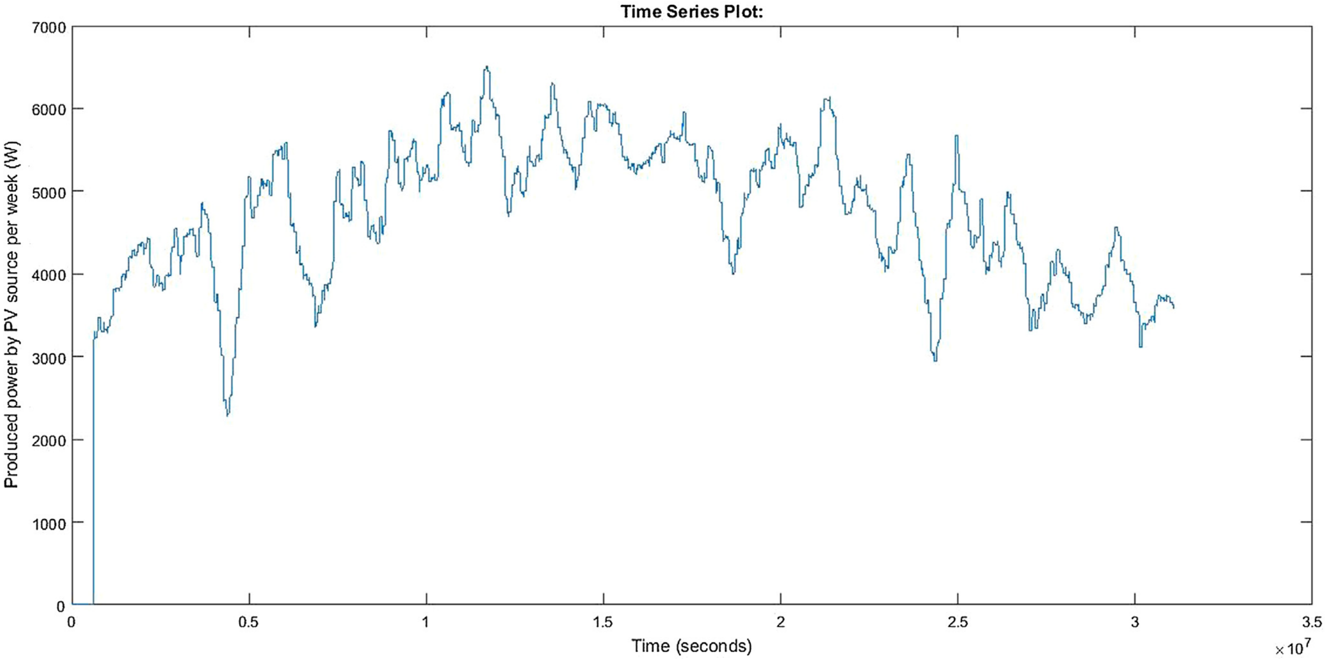


Fig. 5. Power produced by the PV source per week.

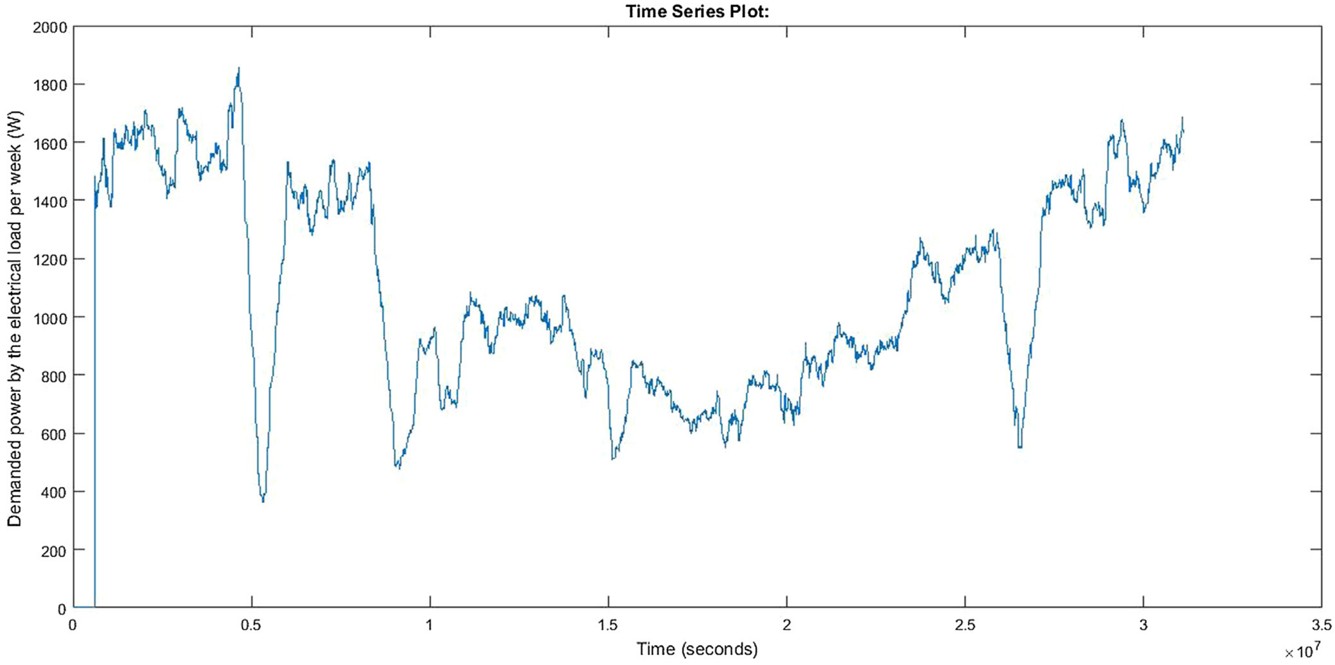


Fig. 6. Demanded power by the load per week.

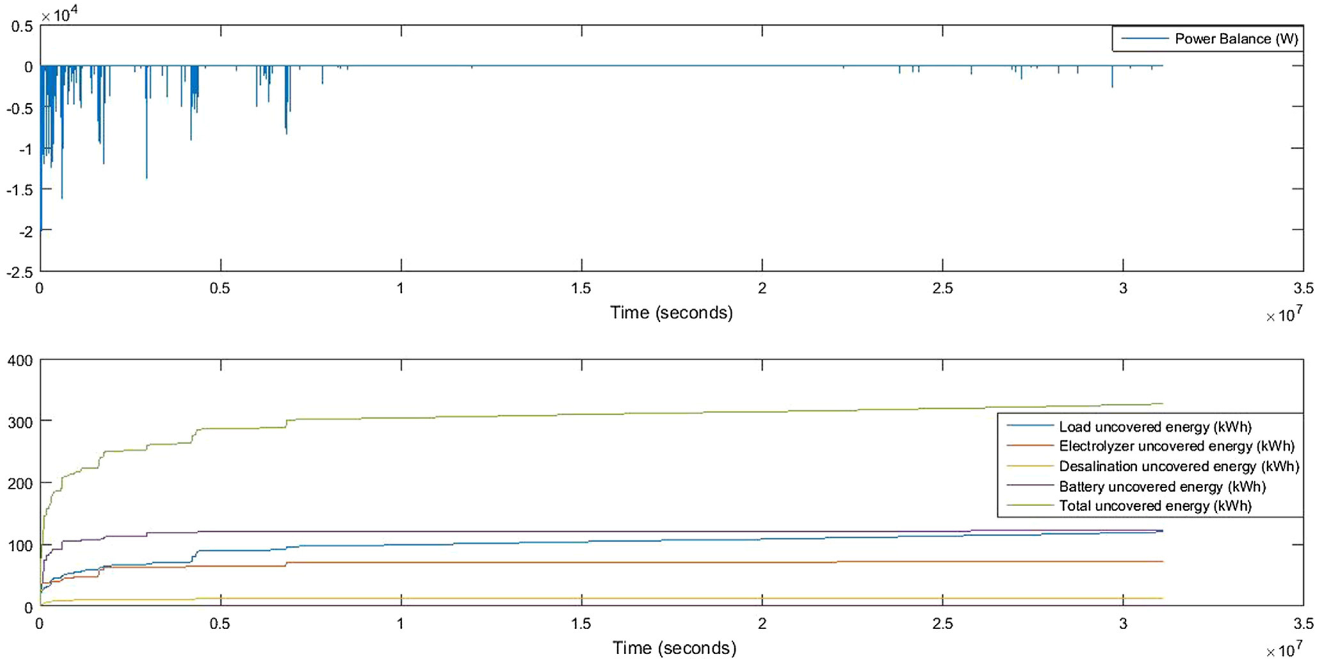


Fig. 7. (a) Power balance between the total consumption and the total production. (b) Total and individual energy demand that has not been covered.

for 26 times the whole year (Fig. 8). A lot of deep discharges occur in the beginning where the battery agent performs exploration (20 deep discharges) and the rest during the winter where the electrical power demand is higher and the Photovoltaic energy production is lower due to less time of sunlight. Deep discharges affect the life time of the battery and have to be minimum. Furthermore, Fig. 9a

Table 1
Efficiency indicators about energy consumption.

	Demanded energy (kWh)	Uncovered energy before convergence (kWh)	Uncovered energy after convergence (kWh)	Total uncovered energy (kWh)	Percentage (%)
Load	9517	99.15 (82.6%)	20.95 (17.4%)	120.1	1.26
Desalination	1939	12.71 (96.7)	0.43 (3.3%)	13.14	0.67
Electrolyzer	3576	71.41 (99.5)	0.36 (0.5%)	71.77	2.01
Battery	6204	121.20 (99.3%)	0.90 (0.7%)	122.1	1.97
Total	21,236	304.47 (93.1%)	22.64 (6.9%)	327.11	1.54

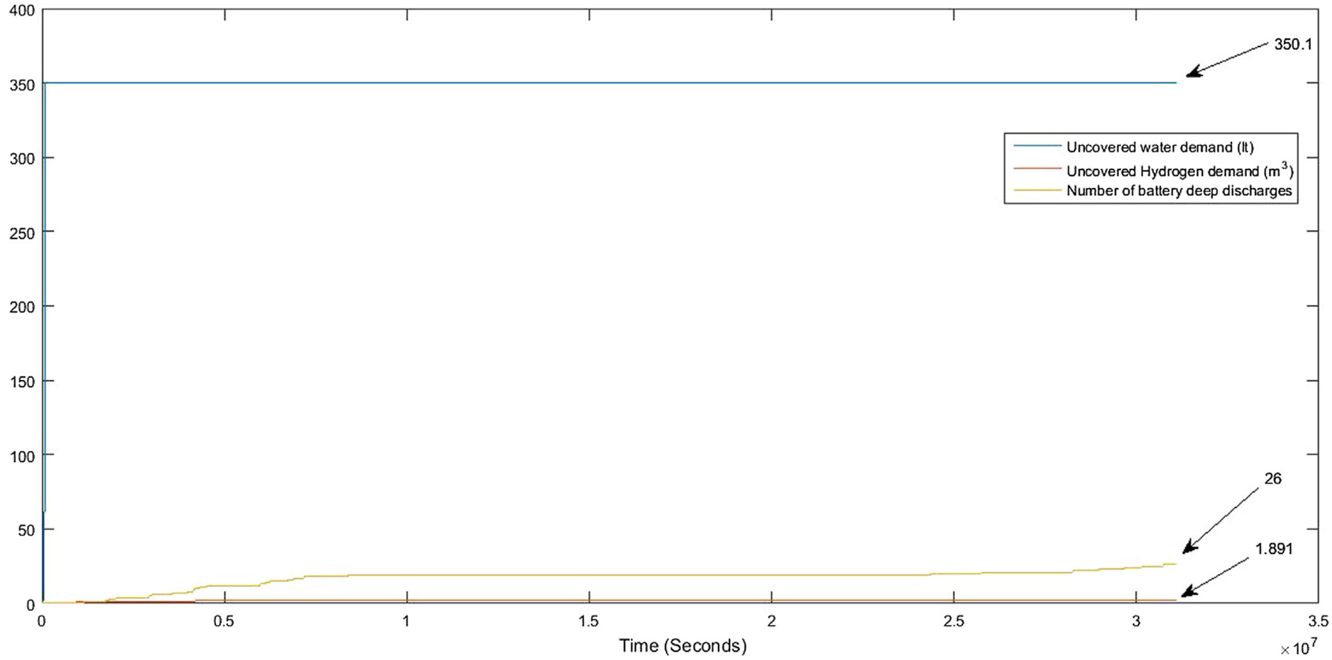


Fig. 8. Uncovered demand of water and Hydrogen, number of Battery deep discharges.

presents the SOC of the battery, the percentage of the stored water and the percentage of the stored Hydrogen for the whole year. Fig. 9b presents the average SOC of the battery, the percentage of the stored water and the percentage of the stored Hydrogen per week. Additional information about the SOC, the water and the hydrogen can be found in Table 2. The uncovered H_2 corresponds to only the 0.41% of the total demanded H_2 and the uncovered amount of the demanded water corresponds to 0.09% of the total demanded amount.

- The delivered power by the diesel generator has to be minimum. The power production and the energy produced by the diesel generator are presented in Fig. 10. The power production is higher in the beginning where the exploration of the agent is in high levels. During the summer the diesel generator does not need to operate at all. In the winter there is are small periods of operation in order to help avoiding the deep discharges of the battery. The total energy production equals to 185.8 and corresponds to only 0.87% of the total produced energy (Table 2).
- The Direct Current (DC) microgrid topology has several advantages concerning stability as there is no need for control of frequency and phase, reactive power and there is no need of synchronization for connection of sources and energy storage systems to the bus [58]. According to [59] "*Power system stability is the ability of an electric power system, for a given initial operating condition, to regain a state of operating equilibrium after being subjected to a physical disturbance, with most system variables bounded so that practically the entire system*

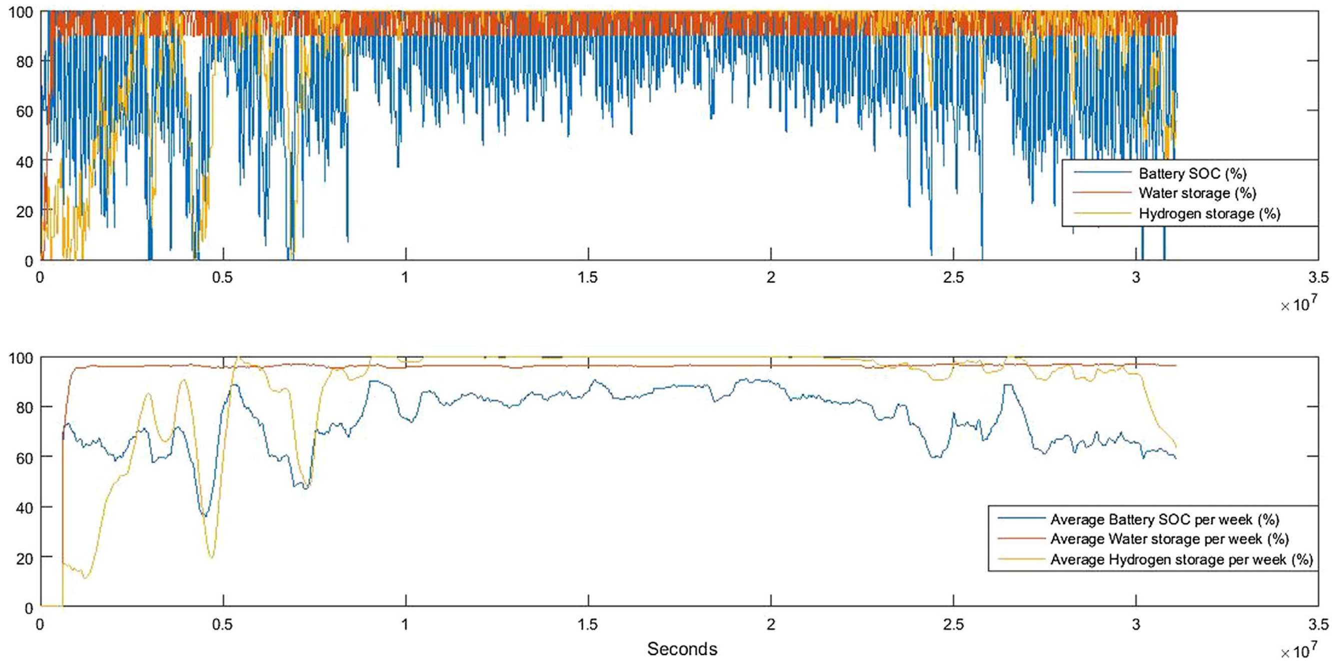


Fig. 9. (a) Battery SOC, wat. (b) Total and individual energy demand that has not been covered.

Table 2
Efficiency indicators about hydrogen, water and battery deep discharges.

Energy produced by diesel generator (kWh)	Number of m^3 of uncovered H_2 demand	Number of liters of uncovered water demand	Number of deep discharges
185.8 (157.1 before convergence and 28.7 after convergence) Cover the 0.87% of the total demanded energy	1.891 (1.891 before convergence and 0.0 after convergence) Does not cover the 0.41% of the total demanded Hydrogen	350.1 (350.1 before convergence and 0.0 after convergence) Does not cover the 0.09% of the total demanded water	26 (19 deep discharges before convergence and only seven after convergence)

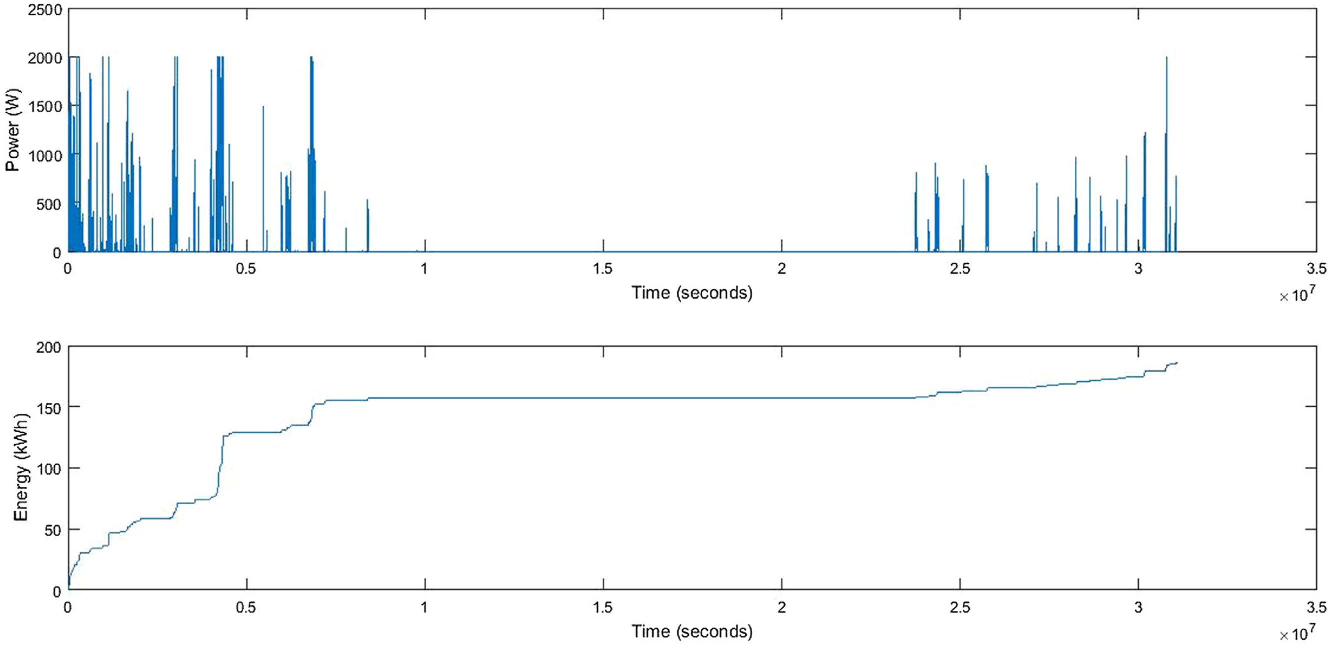


Fig. 10. (a) Power production of diesel generator. (b) Energy produced by diesel generator.

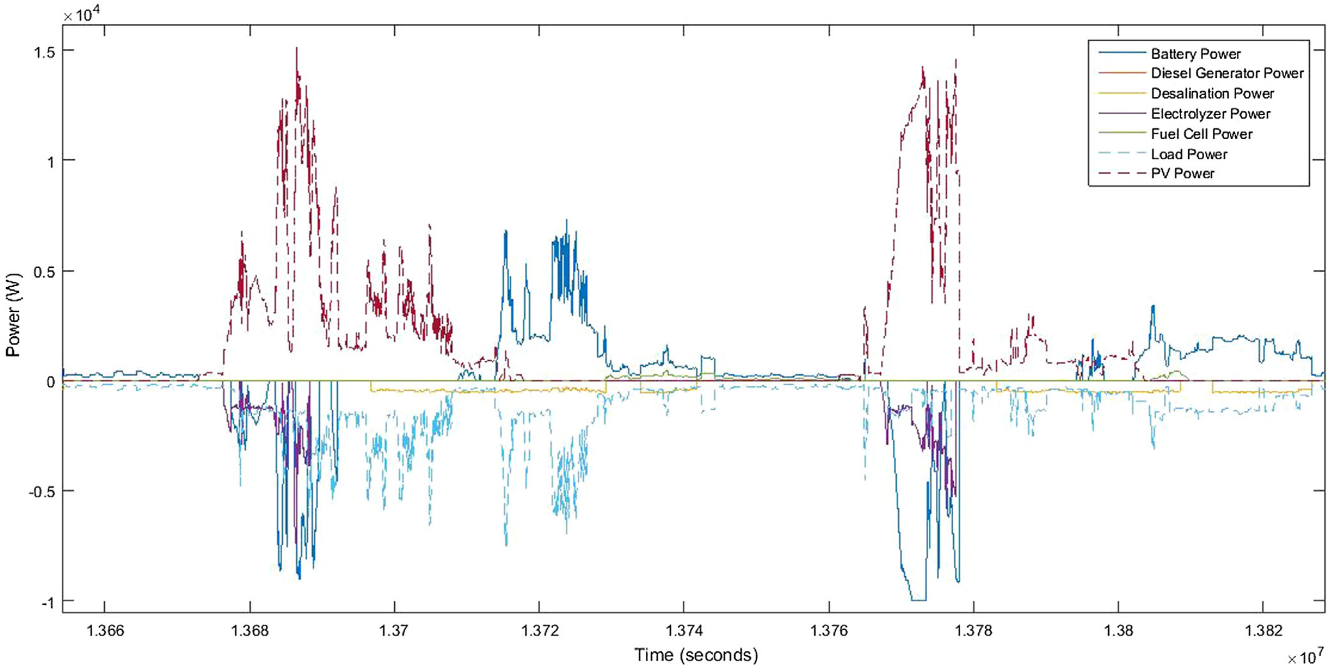


Fig. 11. Power production and consumption of all units for two random consecutive days during summer.

remains intact." The microgrid can be considered as a power system. The stability of the microgrid is evaluated in the scope of power sharing and balance. The PV source delivers the maximum power to the bus and then each agent decides the amount of power injected to the bus. Thus, in the exploration phase, there are many power unbalances and the stability cannot be guaranteed. On the other hand, in the exploitation phase, the agents adequately control the nodes and any power mismatches are restored.

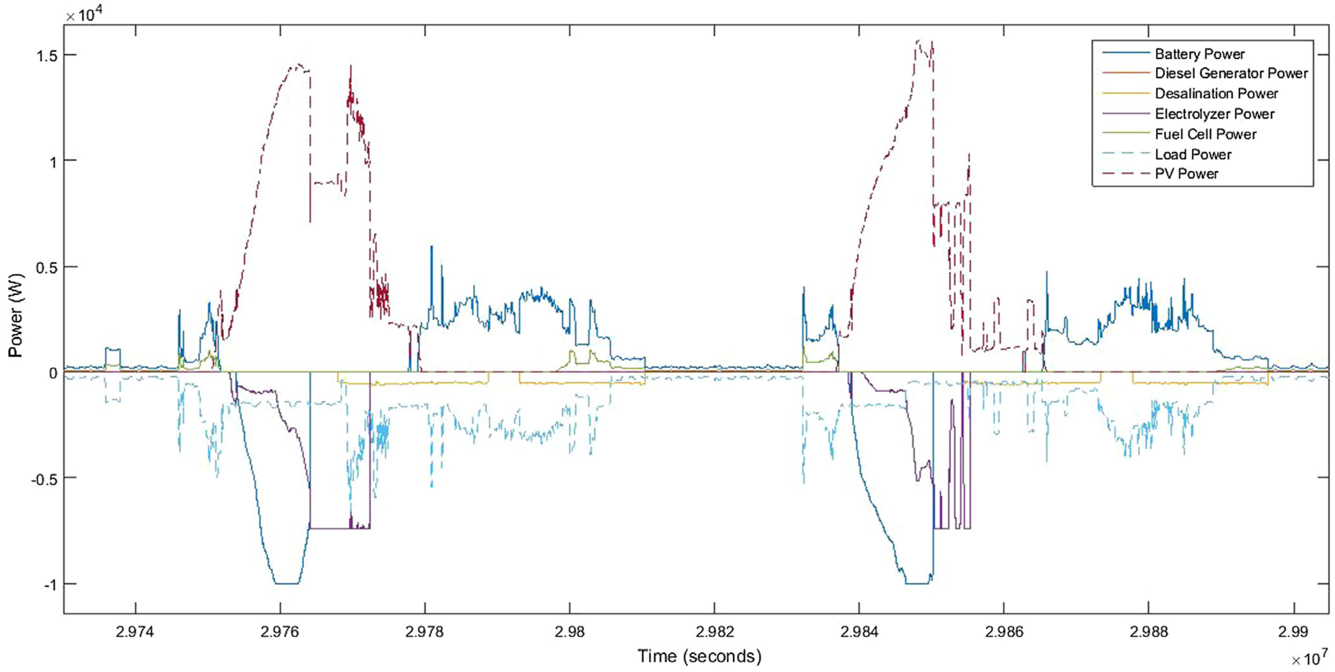


Fig. 12. Power production and consumption of all units for two random consecutive days during winter.

Figs. 11 and 12 present the power production and consumption of all units for two random consecutive days during summer and during winter respectively. While the PV source produces a lot of energy, the demanded power of the load is fully covered by the PV source. Simultaneously, the battery is charging and the electrolyzer consumes energy in order to store Hydrogen. When the power produced by the PV source drops, the electrolyzer stops to consume power and the battery is discharging in order to cover the demanded power arising from the load. The desalination unit operates according to the water needs, the diesel generator does not need to operate at all in both cases. The fuel cell operates when the battery of the SOC drops below 70% and this operation is clearly located during the end of the first day and in the beginning of the second (Fig. 12).

7. Conclusions

This paper demonstrates a MAS to solve the complex problem of energy management in a stand-alone solar microgrid via controlling the energy flow between the microgrid units. In order to reduce the states space and to enhance the learning mechanism, a modified Independent Learners approach is used. This approach is modified by using local rewards and state information that are relevant to each agent, allowing agents to share state variables. Additionally, in order to confront with the continuous state and actions space we introduce fuzzy Q-Learning in each agent. The MAS consists of five agents, three of them have 175 fuzzy states and two of them have 35 fuzzy states. The total amount of fuzzy states which is used equals to 595. Each agent learns through an exploration/exploitation algorithm, converging to a policy very fast, demonstrating good performance. The water tank remains full and only in the beginning, where the agent performs mainly exploration, there is a deficit in the water storage. The battery achieves to provide the appropriate power in the electrical load but some deep discharges are observed (mainly during the exploration phase). The electrolyzer agent manages to keep in high level the hydrogen storage and manages to serve the hydrogen's needs of the fuel cell. The fuel cell agent and the diesel generator agent act ancillary to the battery in order to avoid the deep discharges of the battery. The deep discharges of the battery are limited to 26 for the time period of a year, while these are very few when the agent has been trained and converged to a policy.

Despite the fact that the independent learners approach cannot guarantee convergence (the main drawback of this method, mentioned in Section 3.3), in this case study, the simulation results indicate the algorithm convergence; highlight both the individual performance of the agents and the total performance of the providing MAS. The convergence may be slow, but after that the percentage of time agents explore is low, in order for them to detect any changes in the topology, or the units of the microgrid and re-adapt the policy. This makes possible to apply the trained algorithm in any similar system and avoid the initial intense exploration: However the tolerance to any such changes and the capability for adaptation is something to be further investigated. Having said that, we must point out that the proposed MAS can be trained offline and then can be applied to a real system. In any case optimality can-not be guaranteed and thus solutions are considered in the general case suboptimal.

8. Future work

In future, it is our aim to perform comparative experiments with different MAS approaches for both grid-connected and island-mode microgrids. Furthermore, we plan to apply these techniques to real-world settings where a microgrid contains more units, such as a wind turbines and hybrid electrical vehicles.

References

[1] Zeng J, Wu J, Jun-feng L. An agent-based approach to renewable energy management in eco-building. Sustainable energy technologies, 2008. ICSET 2008. IEEE international conference on. 2008. p. 46–50.

[2] Skarvelis-Kazakos S, Papadopoulos P, Undac IG, Gormana T, Belaidi A, Zigan S. Multiple energy carrier optimisation with intelligent agents. Appl Energy 2016;167(1):323–35.

[3] Sechilariu M, Wang BC, Locment F, Jouglet A. DC microgrid power flow optimization by multi-layer supervision control. Design and experimental validation. Energy Convers Manage 2014;82:1–10. http://dx.doi.org/10.1016/j.enconman.2014.03.010.

[4] Dimeas AL, Hatziargyriou ND. Operation of a multiagent system for microgrid control. IEEE Trans Power Syst 2005;20(3):1447–55.

[5] Smith M. DOE microgrid initiative overview. Paper presented at conference on 2012 DOE microgrid workshop, Chicago, Illinois. 2012.

[6] El-Sharafy MZ, Farag HEZ. Back-feed power restoration using distributed constraint optimization in smart distribution grids clustered into microgrids. Appl Energy 2017;206(15):1102–17.

[7] Coelho VN, Cohen MW, Coelho IM, Liu N, Guimarães FG. Multi-agent systems applied for energy systems integration: state-of-the-art applications and trends in microgrids. Appl Energy 2017;187:820–32.

[8] Kofinas P, Vouros G, Dounis AI. Energy management in solar microgrid via reinforcement learning. ACM international conference proceeding series 18-20-May-2016, 9th Hellenic conference on artificial intelligence, SETN 2016, Thessaloniki, Greece. 2016.

[9] Hossain E, Kabalci E, Bayindir Ramazan, Pereza R. Microgrid testbeds around the world: state of art. Energy Convers Manage 2014;86:132–53.

[10] Farhangi H. Intelligent micro grid research at BCIT. Electric power conference, 2008. IEEE Canada, 6–7 October 2008. 2008. p. 1–7.

[11] Jin Xiaolong, Jianzhong Wu, Yunfei Mu, Wang Mingshen, Xiandong Xu, Jia Hongjie. Hierarchical microgrid energy management in an office building. Appl Energy 2017;208:480–94.

[12] Mahmoud MS, Azher Hussain S, Abido MA. Modeling and control of microgrid: an overview. J Franklin Inst 2014;351(5):2822–59.

[13] Tsikalakis AG, Hatziargyriou ND. Centralized control for optimizing microgrids operation. IEEE Trans Energy Convers 2008;23(1):241–8.

[14] Hong T, Bian T, de León F. Supplementary damping controller of grid connected dc micro-grids based on Q-learning. IEEE power and energy society general meeting (PESGM), Boston, MA. 2016. p. 1–5.

[15] Li Q, Chen F, Chen M, Guerrero J, Abbott D. Agent-based decentralized control method for islanded microgrids. IEEE Trans Smart Grid 2016;7(2).

[16] Shi B, Liu J. Decentralized control and fair load-shedding compensations to prevent cascading failures in a smart grid. Int J Electr Power Energy Syst 2015;67:582–90.

[17] Sycara K. Multiagent systems. AI Mag 1998;19(2):79–92.

[18] Kok Jelle R, Vlassis N. Collaborative multiagent reinforcement learning by payoff propagation. J Mach Learn Res 2006;7:1789–828.

[19] Dounis AI, Caraiscos C. Advanced control systems engineering for energy and comfort management in a building environment – a review. Renew Sustain Energy Rev 2009;13(6-7):1246–61.

[20] Anvari-Moghaddam Amjad, Rahimi-Kian Ashkan, Mirian Maryam S, Guerrero Josep M. A multi-agent based energy management solution for integrated buildings and microgrid system. Appl Energy 2017;203:41–56.

[21] Wang Z, Wang L, Dounis AI, Yang R. Multi-agent control system with information fusion based comfort model for smart buildings. Appl Energy 2012;99:247–54.

[22] Rahman MS, Oo AMT. Distributed multi-agent based coordinated power management and control strategy for microgrids with distributed energy resources. Energy Convers Manage 2017;139:20–32.

[23] Xydas E, Marmaras C, Cipcigan LM. A multi-agent based scheduling algorithm for adaptive electric vehicles charging. Appl Energy 2016;177(1):354–65.

[24] Kim H-M, Lim Y, Kinoshita T. An intelligent multiagent system for autonomous microgrid operation. Energies 2012;5:3347–62.

[25] Foo YSE, Gooi HB, Chen SX. Multi-agent system for distributed management of microgrids. IEEE Trans Power Syst 2014;30:24–34.

[26] Ma L, Liu N, Zhang JH, Tushar W, Yuen C. Energy management for joint operation of CHP and PV prosumers inside a grid-connected microgrid: a game theoretic approach. IEEE Trans Ind Inform 2016:1930–42.

[27] Pipattanasomporn M, Feroze H, Rahman S. Multi-agent systems in a distributed smart grid: design and implementation. Proc. IEEE/PES power syst. conf. expo. (PSCE), Seattle, WA, USA, Mar. 2009. 2009. p. 1–8.

[28] Chung I-Y, Yoo C-H, Oh S-J. Distributed intelligent microgrid control using multi-agent systems. Engineering 2013;5:1–6.

[29] Koohi-Kamali S, Rahim NA. Coordinated control of smart microgrid during and after islanding operation to prevent under frequency load shedding using energy storage system. Energy Convers Manage 2016;127:623–46.

[30] Zhao Bo, Xue Meidong, Zhang Xuesong, Wang Caisheng, Zhao Junhui. An MAS based energy management system for a stand-alone microgrid at high altitude. Appl Energy 2015;143:251–61.

[31] Khan MRB, Jidin R, Pasupuleti J. Multi-agent based distributed control architecture for microgrid energy management and optimization. Energy Convers Manage 2016;112:288–307.

[32] Logenthiran T, Srinivasan D, Khambadkone AM, Aung HN. Multiagent system for real-time operation of a microgrid in real-time digital simulator. IEEE Trans Smart Grid 2012;3(2).

[33] Kyriakarakos G, Piromalis D, Dounis AI, Arvanitis KG, Papadakis G. Intelligent demand side energy management system for autonomous polygeneration microgrids. Appl Energy 2013;103:39–51.

[34] Shirzeh H, Naghdy F, Ciufo P, Ros M. Balancing energy in the smart grid using distributed value function (DVF). IEEE Trans Smart Grid 2015;6(2).

[35] Makrygiorgou Despoina I, Alexandridis Antonio T. Distributed stabilizing modular control for stand-alone microgrids. Appl Energy 2018;210:925–35.

[36] Riverso S, Tuccib M, Vasquez JC, Guerrero JM, Ferrari-Trecate G. Stabilizing plug-and-play regulators and secondary coordinated control for AC islanded microgrids with bus-connected topology. Appl Energy 2018;210:914–24.

[37] Kofinas P, Vouros G, Dounis AI. Energy management in solar microgrid via reinforcement learning using fuzzy reward. Adv Build Energy Res 2017;1–19.

[38] Tsoukalas L, Uhring R. Fuzzy and neural approaches in engineering. MATLAB Suppl 1997.

[39] Zadeh LA. Fuzzy sets. Inf Control 1965;8(3):338–53.

[40] Wang Li-Xin. A course in fuzzy systems and control. Prentice Hall PTR; 1997.

[41] Jang J, Sun C, Mizutani E. Neuro-fuzzy and soft computing. Prentice Hall; 1996.

[42] Russel S, Norving P. Artificial intelligence: a modern approach. Upper Saddle River (NJ): Prentice Hall; 1995.

[43] Watkins CJCH. Learning from delayed reinforcement signals [PhD Thesis]. England: University of Cambridge; 1989.

[44] Reinforcement learning: an introduction. Richard Sutton and Andrew Barto. MIT Press; 1998.

[45] Vincent François-Lavet, Fonteneau Raphael, Ernst Damien. How to discount deep reinforcement learning: towards new dynamic strategies. NIPS, Deep RL workshop; 2015

[46] van Hasselt Hado. Reinforcement learning in continuous state and action spaces. Reinforcement learning: state of the art. Springer; 2012. p. 207–51.

[47] Castro JL. Fuzzy logic controllers are universal approximators. IEEE Trans SMC 1995;25(4).

[48] Glorennec PY, Jouffe L. Fuzzy Q-learning. Proceedings of 6th international fuzzy systems conference. 1997. p. 659–62.

[49] Puterman ML. Markov decision processes: discrete stochastic dynamic programming. New York: Wiley; 1994.

[50] Guestrin C, Koller D, Parr R. Multiagent planning with factored MDPs. Advances in neural information processing systems (NIPS). The MIT Press; 2002.

[51] Guestrin C, Lagoudakis M, Parr R. Coordinated reinforcement learning. Proceedings of international conference on machine learning (ICML), Sydney, Australia. 2002.

[52] Schneider J, Wong W-K, Moore A, Riedmiller M. Distributed value functions. Proceedings of international conference on machine learning (ICML), Bled, Slovenia. 1999.

[53] Claus C, Boutilier C. The dynamics of reinforcement learning in cooperative multiagent systems. Proceedings of the national conference on artificial intelligence (AAAI), Madison, WI. 1998.

[54] Laurent GJ, Matignon L, Fort-Piat NL. The world of independent learners is not markovian. KES J 2011;15(1):55–64.

[55] Busoniu L, Schutter BD, Babuska R. Multiagent reinforcement learning with adaptive state focus. BNAIC. 2005. p. 35–42.

[56] Thirugnanam K, Kerk SK, Yuen C, Liu N, Zhang M. Energy management for renewable micro-grid in reducing diesel generators usage with multiple types of battery. IEEE TIE; 2018.

[57] Kofinas P, Dounis AI, Mohamed ES, Papadakis G. Adaptive neuro-fuzzy model for renewable energy powered desalination plant. Desalinat Water Treat, 65. 2017. p. 67–78.

[58] Shafiee Qobad, Dragicevic Tomislav, Vasquez Juan C, Guerrero Josep M. Modeling, stability analysis and active stabilization of multiple DC-microgrid clusters. Proc. IEEE international energy conference (EnergyCon'14). 2014.

[59] Kundur P, Paserba J, Ajjarapu V, Andersson G, Bose A, Canizares C, et al. Definition and classification of power system stability IEEE/CIGRE joint task force on stability terms and definitions. IEEE Trans Power Syst 2004;19(3):1387–401.