Virtual Power Plant Day Ahead Energy Unit Commitment

Andres F. Ramirez
University of Los Andes, Bogota, Colombia
af.ramirez12@uniandes.edu.co

Alberto J. Lamadrid
Lehigh University, PA, US
ajlamadrid@lehigh.edu

Carlos F. Valencia
University of Los Andes, Bogota, Colombia
cf.valencia@uniandes.edu.co

Abstract

In this article we present a model for the interaction of distributed energy resources (DER) with the electricity system, using reinforcement learning. Our method relaxes the requirements for information necessary to train and engage in Pareto improving trading, and can directly incorporate the inherent intermittency of variable renewable energy sources. The distributed resources include consumers of electricity, energy storage systems, and variable renewable energy. We modify the algorithms to improve the scheduling of the resources. In our empirical application, we use data from Colombia subject to large variability due to El Niño Southern Oscillation and illustrate the use of the model under large variations in the data used to train the model.

1. Introduction

Many countries, including Latin American ones such as Colombia, have accelerated their efforts to increase the amount and diversity of their electricity generation capabilities. For instance, the generation portfolio of Colombia primarily consists of hydro-electric power. This energy source accounts for nearly 70% of the total energy production of the country [1]. This lack of diversity responds to the fact that Colombia is a country rich in water resources; nonetheless, the dependence on hydroelectric power plants can have dire consequences. In particular, weather phenomena like El Niño Southern Oscillation (ENSO) often create emergencies across the country generating drastic droughts that require production of energy with non-renewable energy sources, such as coal or natural gas. To respond to these issues in 2019 the Colombian government announced the construction of *Parque Solar Castilla*, a solar photovoltaic farm with an installed capacity of 21 megawatts(MW). Moreover, the government announced two more projects, with a projected installed capacity of 80MW and 50MW.

The integration of intermittent and variable renewable energy sources (VRES) poses challenges in terms of the balance of supply and demand in the electricity system. Currently, most systems have limited responsiveness from the demand side. In fact, the price-inelastic demand requires additional balancing services, driven mostly by the supply side. Moreover, several challenges arise when estimating the effectiveness of a protect, such as the need to understand the weather variables that directly affect energy production.

The increased participation of VRES in the generation portfolios has nurtured interest in Energy Storage Systems (ESS). The research in the development of technologies, models, and policies for coupled ESS is an active area [see e.g., 2].

A fundamental operational problem to solve is the feasible management of the resources to attain the objectives desired, including the maximization of social welfare or minimization of production costs, given the uncertainties present. Additionally, the institutional arrangements affect the feasibility and *implementability* of the interaction schemes.

Our paper presents a model for the integration of VRES with the electricity system, using a combination of state-of-the-art methodologies. The DER include consumers of electricity, ESS and VRES. We propose the integration of VRES in the context of a price taker virtual power plant (VPP) modeling the inherent uncertainty and using a novel methodology for optimal decision policy making. The novelty of the manuscript is the combination and application of (i) An implementable method for resource sharing that generates optimal day ahead scheduling policies using VRES as the main energy production source. In addition, allowing and integration of profile generator units, optimal scheduling units, among others on a scheme where a *prosumer* interacts as peers in a smart electrical grid (see e.g. Figure 1). (ii) A flexible Deep Reinforcement Learning (DRL) model applying Temporal Difference (TD) learning methods

URI: https://hdl.handle.net/10125/79756
978-0-9981331-5-7

with experience replay for validating and committing day ahead decisions on a real time market. (Algorithm 2) (iii) Integration of two hierarchical Markov decision processes where day-ahead actions are evaluated by the real-time deep reinforcement learning agent using the constrained cross entropy approach (Algorithm 1). (iv) Better representation of the stochastic nature of the problem VRES energy production uncertainty (solar and wind) using state-of-the-art simulation methods; in addition, we compare our results with an naive deterministic agent showcasing how a more deterministic approach can get heavily affected by extreme scenarios. Finally, (v) Reduction on the amount of system information needed to optimally control distributed resources in comparison to robust optimization approaches allowing more flexibility to changing environments as well.

2. Methodology

The virtual power plant (VPP) day-ahead unit commitment methodology can be summarized in two main components: (i) a deep reinforcement learning methodology using experience replay based on expected State–action–reward–state–action (SARSA) paired with a constrained cross entropy method to solve the optimal control problem (Section 2.9), and (ii) a multivariate copula autoregressive algorithm for simulating solar irradiance, air temperature and wind speed (Section 2.13). The combination of these two modules for solving the optimal control problem has two main advantages. First, it relaxes the requirement of previous knowledge of the system that is inherent to methodologies such as robust control. These requirements limit the available configurations of the system and generate *ad hoc* solutions for specific case studies or system designs. The robust controller requires a vast *a priori* comprehension of the system dynamics and this knowledge cannot be changed. Conversely, the reinforcement learning controller uses limited previous knowledge of the system but is more flexible and capable to adapt to changing environments of system configurations to find optimal control schemes [3]. Second, we seek to incorporate synthetic data in order to emulate the natural variability of VRES in the optimal control problem. For this, we take advantage of a methodology for simulation of time series of natural variables (such as wind speed and solar irradiance) that works based on a copula autoregressive methodology. This allows a more accurate simulation of the phenomena to reproduce the variability of the process and reduce the uncertainty related to energy production.

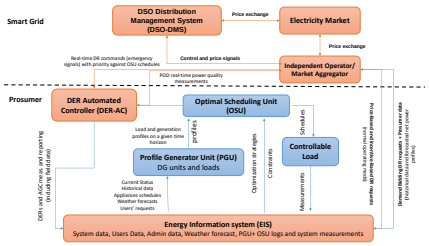


Figure 1: Scheme for the integration of the proposed virtual power plant (VPP) optimal control methodology on a service-oriented multi agent scheme

2.1. Problem Formulation

We model a theoretical VPP system with a hierarchical sequential Markov Decision Process (MDP). This process has two main components that represent the usual decision levels in current systems: the day ahead (DA) decisions process (DA-MDP agent) describes the actions required for the day-ahead energy commitment to be evaluated. The real-time (RT) decision process (RT-MDP agent) describes the state of the VPP, its performance, and the hourly basis decisions that seek to minimize the cost and maximize the reliability of the system. DA decisions are taken by the DA agent in order to maximize the day ahead energy commitment of the VPP. However, the reliability of the energy commitment of the DA agent can only be assessed once the uncertainty has been cleared in RT, and is dependent on the internal system dynamics governed by the decisions of the RT agent. This results in a complex dependence between the RT and DA agents with system reliability and cost-effectiveness. Although we are using two agents, we are not in the spectrum of a multi-agent RL (MARL) problem, because the agents are not sharing the same environment, their actions and task are different from each other and they are not interacting simultaneously [4; 5; 6].

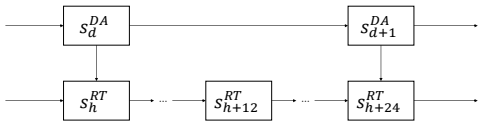


Figure 2: Day ahead (DA) and Real Time (RT) hierarchical MDPs. The RT process evaluates the decisions taken in the DA process.

2.2. Real Time MDP

The RT-MDP is a tuple $(S^{RT}; A^{RT}; P^{RT}; R^{RT})$ that represents the real-time reliability and cost-minimizing control process. The time index for this process is t, denoting intra-day hourly time steps. In RT, the operator of the VPP system may choose preventive actions at each time step, trying to immunize the system against system and weather variability (risk) by attempting to avoid unreliable states. To model this decision process in each interval, the agent observes the current state $s_t^{RT} \in S^{RT}$ (i.e. the state of charge of the battery, the realized energy commitment, local demand, and the solar and wind energy production, etc. for this interval) and chooses an action $a_t^{RT} \in \mathbb{A}^{RT}$. Following the agent's action, the real-time reward $r_t^{RT}(S_t^{RT}, a^{RT})$, representing the cost to guarantee the fulfillment of the energy commitment and the system's reliability, can be calculated, and a transition to s_{t+1}^{RT} occurs, governed by the respective policy $P(s_{t+1}^{RT}|s^{RT}, a^{RT})$.

2.3. Real time State-space

We define a RT state s_t^{RT} to be the tuple (α,β,γ,δ,λ,ϵ,ζ,η,θ), where:

- α Is the state of charge of the battery in the current time step $SOC_{t+1} = SOC_t + \text{Solar}_t + \text{Wind}_t + \text{MarketPurchase}_t$ where Solar and Wind is the amount of energy produced by each one of the renewable sources respectively and Market Purchase is the amount of energy acquired from the market at time t in this day.
- β Is the hour of the day in the current time step
- γ Is the day of the month in the current time step
- δ Is the month of the year in the current time step
- λ Is a vector with 24 positions for simulated solar irradiance and wind speed using the copula autoregressive algorithm 2.13
- ϵ Is the price of energy (e.g. \$/MWh or \$/kWh) taken from the market for the current time step

  We assume the VPP agent is a price taker in the energy market. We use information for the Colombian energy prices provided by XM, the system operator that manages and operates the electricity market in Colombia [7]
- ζ Is the amount of energy commitment (e.g. MWh or kWh) for the current time step (e.g., hour)
- η Is the amount of energy produced with the renewable energy sources for the current time step
- θ Is the amount of energy purchased (e.g. MWh or kWh) to the market for the current time step

2.4. Real time Action-Space

An action a_t^{RT} in RT attempts to achieve improved reliability of the system by increasing the amount of available energy in periods of time of high demand and low renewable energy production. The action the RT agent might take at each step is the amount of energy to buy from the market. i.e., increasing the state of charge (SOC) for the next hour to be able to better respond to the energy commitment previously chosen by the DA agent:

$$S_t^{RT} \xrightarrow{a_t^{RT}} S_{t+1}^{RT} = (\alpha, \beta, \gamma, \delta, \epsilon, \zeta, \eta, \theta + \Delta\theta)$$

The actions selected by the agent are sampled from the interval $[0\%, 100\%]$. That is, the agent could buy 100% of the current battery capacity to use in the next time steps.

2.5. Real time Transition Kernel

The RT transition kernel is conditioned on the last RT state and action, and on the corresponding last DA decision taken to determine the energy commitment in that specific time t of the day:

$$S_{t+1}^{RT} = f(S_t^{RT}, a_t^{RT}, a_{t_d}^{DA}).$$

The dependence between RT and DA states is expressed using the following sets of equations. The RT demand process is based on the DA energy commitment:

$$d_t^{RT} = d_t^{DA} + \delta_t, \tag{1a}$$

$$\delta_{t+1} = \delta_t + \epsilon_t. \tag{1b}$$

where d_t^{RT} is the RT demand vector at time t, d_t^{DA} is the DA energy commitment vector for time t of the day; finally, ϵ_t is Gaussian noise and δ_t is the autoregressive random bias parameter [8].

2.6. Real time Reward

We choose the RT reward to be the profit received by the agent at time t of the day. This profit is calculated as the difference between the amount of revenue by selling stored and produced energy minus the cost related to storing that energy, production costs, and cost related to penalties for overproducing energy or for not having enough energy to dispatch and respond to the energy commitment of the DA agent. To calculate the profit of the system in every time step we evaluate three possible

cases. The first case is when the energy available in the VPP is higher than the battery capacity of the system. This over-production generates a penalty equal to the amount of energy over the installed battery capacity that is then added to the total cost of the agent step. The second case is when the available energy is less than the energy commitment of the DA agent. In this case, the agent needs to pay a penalty for the amount of energy that was not produced. Finally, if the amount of energy available in the system is less than the capacity of the system and more than the energy commitment, the agent gains the amount of energy sold to the market. With these cases, we calculate the profit at each time t of the day for every step that the agents take.

2.7. Hierarchical MDP sequential Policy Improvement algorithm

We use an algorithm for jointly learning the RT cost-effective value function and simultaneously search for an optimal DA policy. This follows a similar approach to the one conducted in [9; 10; 11]. This is a feedback loop, and the main idea of the algorithm is that the policy improvement of the DA MDP is based on the RT MDP value function. Meanwhile, the RT MDP value function is being affected by the day ahead decisions (energy commitment) of the DA MDP. We use a deep reinforcement learning algorithm to assess the RT reward on each step (system cost-effective profit) and the cross entropy method [12; 13] adding constraints to the selection of elite paths as presented by [14] for improving the DA policy.

2.8. Day Ahead Policy Improvement and Comparison

For the VPP design we assume that the operator will use the installed energy production. Thus the DA agent decision is similar to the work proposed by [9], and consists of a parametric DA policy as $\pi(s^{DA};\psi) = \arg\max_{a^{DA}\in\mathbb{A}^{DA}}\Phi(s^{DA})$

A comparison between different DA policies π_i is done by calculating the empirical expected value of the RT value function,

$$\mathbb{E}\nu^{\pi_i} \approx \frac{1}{|\mathbb{S}^{RT}_{paths}|}\sum_{s\in\mathbb{S}^{RT}_{paths}} \nu^{\pi_i}(s), \tag{2}$$

using a set of representative RT initial states $\mathbb{S}^{RT}_{paths}$. This set is composed of the full history of all RT states visited during the current Hierarchical MDP sequential policy improvement algorithm iteration. This procedure allows the DA agent to compare different policies using many probable states that are being sampled from the distribution $P^{(k)}_{\psi}$.

DA policy improvement is achieved using the constrained cross entropy method. In this method, initial policies are sampled from a distribution $P^{(0)}_{\psi}$. After sampling, in each iteration, k policy parameters are drawn from $P^{(k)}_{\psi}$, and their top percentile, according to the RT value following the constraint function H, is used to update $P^{(k+1)}_{\psi}$. The distribution $P^{(k)}_{\psi}$ is a Gaussian mixture with means set to ψ^{k+1}, which is the energy load that belongs to the top percentile of all the paths. The convergence criterion we use in our instances is:

$$\frac{1}{N_{top}}\sum_{i=1}^{N_{top}}(\hat{\nu_i}^{(k)} - \hat{\nu_i}^{(k-1)})^2 < \epsilon,$$

Algorithm 1: Hierarchical MDP sequential policy improvement algorithm

Input: initial distribution $P^{(0)}_{\psi}$ for DA policy parameters and a constraint function H, upper bound d and convergence limit ϵ

Output: optimal DA policy $\pi(\psi^*)$

Initialize $\mathbb{S}^{RT}_{paths} = \emptyset, k = 0$

Convergence measure:

$$\frac{1}{N_{top}}\sum_{i=1}^{N_{top}}(\hat{\nu_i}^{(k)} - \hat{\nu_i}^{(k-1)})^2 < \epsilon$$

while *Convergence measure* **do**
- **for** $i \leftarrow 1$ **to** N **do**
  - Get load from distribution $\psi_i \sim P^{(k)}_{\psi}$
  - Sample N_{paths} from DA policy $\pi_i = \pi(\psi_i)$
  - Approximate $\nu^{\pi_i}\ \forall\ N_{paths}$
  - Approximate $H(\pi_i)$ using ν^{π_i} approximation $\forall\ N_{paths}$
  - Save DRL trajectories to $\mathbb{S}^{RT}_{paths}$
- set $\hat{\nu_i}^{(k)} = \frac{1}{|\mathbb{S}^{RT}_{paths}|}\sum_{s\in\mathbb{S}^{RT}_{paths}}\nu^{\pi_i}, \forall i \in |N|$
- **if** $H(\pi_i) \leq d$ **then**
  - sort $\pi_i | H(\pi_i) \leq d$ in descending order with respect to $\hat{\nu_i}$
- **else**
  - **end if**
- Update $P^{(k)}_{\psi}$ using the respective ψ_i from the top percentile of π_i
- k+=1

where N_{top} is the number of the top-percentile values. By using the constrained cross entropy method, we avoid using gradient-based optimization which may be difficult to compute in our case due to the interaction needed between the agents. Nonetheless, we are not in the spectrum of a multi-agent RL (MARL) [see e.g., 4; 5; 6]

2.9. Real Time Value Function Approximation

The RT agent environment is based on a virtual power plant (VPP) environment. Since a precise model of the system is not available and detailed restrictions, information, or characteristics of the system are scarce, it is preferred to estimate action values $q_*(s,a)$ (the values of state-action pairs) rather than state values $v_*(s)$. Accordingly, a primary goal for the reinforcement learning method used is to estimate the optimal action-value function q^*. The policy evaluation problem for action values is to estimate $q_\pi(s,a)$, the expected return when starting in-state s, taking action a, and thereafter following policy π [15].

$$v_*(s) = \max_a \mathbb{E}[R_{t+1} + \gamma v_*(S_{t+1})|S_t = s, A_t = a]$$

$$= \max_a \sum_{s',r} p(s', r|s, a)[r+\gamma v_*(s')|S_t = s, A_t = a], \text{or,} \tag{3}$$

$$q_*(s,a) = \mathbb{E}[R_{t+1}+\gamma \max_{a'} q_*(S_{t+1}, a')|S_t = s, A_t = a]$$

$$= \sum_{s',r} p(s', r|s, a)[r + \gamma \max_{a'} q_*(s', a')]. \tag{4}$$

In the following sections we give a detailed explanation of the methodology used for Approximating ν^{π_i} (Algorithm 2) using DRL with expected SARSA. We begin by giving details on why we used Temporal difference (TD) learning in the form of Expected SARSA, after that we give the benefits of using experience replay in this setup finishing with the action selection procedures chosen and the Deep Expected SARSA algorithm used.

2.10. Experience Replay

Reinforcement learning methods have the restriction that the learning and updating procedure must follow a sequential order. Every sample from the interaction of the agent with the environment generates one update to the value functions, thus making the learning process inefficient. Moreover, the approximation of the value function using a Neural Network can be affected since the algorithm is learning from experience tuples in sequential order that can generate correlation problems.

To avoid this problem we use experience replay for generating a training data set from a buffer of sample data from the environment. This technique is called replay buffer or experience buffer [16]. The replay buffer consists of experience tuples ($\mathcal{S}$, $\mathcal{A}$, $\mathcal{R}$, $\mathcal{S}+$). As well as minimizing undesired correlations, experience replay allows the agent to learn from individual tuples multiple times, and recall rare occurrences using the experience of the environment in a more efficient way.

2.11. Soft-max policy action selection

We select the soft-max algorithm for action selection. The soft-max policy parametrization explores according to the action-values. That is, an action with a moderate value has a higher chance of getting selected compared to an action with a lower value [17]. We use this parametrization given that the estimation of the approximate policy can approach a deterministic policy, whereas with ϵ-greedy action selection over action values there is always an ϵ probability of selecting a random action. The VPP control problem has high variability and uncertainty due to the dependence on weather variables. Accordingly, the best approximate policy for this particular problem may be stochastic. According to the probability of selecting each action, the soft-max policy dependent on the state s and action a where τ is the temperature parameter which controls how much the agent focuses on the highest valued actions. The lower the temperature, the more the agent selects the greedy action. When the temperature is high, the agent selects among actions in a more random way. The softmax policy exponentiates action values; this characteristic could diverge and generate large action value functions. As a consequence we implemented the soft-max policy in a numerically stable way, subtracting the maximum action-value from the action-values as follows:

$$Pr(A_t = a|S_t = s) \doteq \frac{e^{Q(s,a)/\tau - max_c Q(s,c)/\tau}}{\sum_{b \in A} e^{Q(s,b)/\tau - max_c Q(s,c)/\tau}}. \tag{5}$$

2.12. Deep Expected SARSA learning

A virtual power plant environment (VPP) can be regarded as an MDP. We use Expected SARSA integrated into a DRL framework to solve the optimal control problem. The current state-action value is $Q(s,a)$. We have an action-value function represented as a neural network, $Q_t(s,a)$. We update our

action-value function and get a new one that we can use at the next time-step. We will get this $Q_{t+1}(s,a)$ using multiple replay steps that each result in an intermediate action-value function $Q_{t+1}(s,a)^i$ where i indexes the replay step being evaluated. For an in-depth analysis of the solution of deep linear neural networks, the reader is referred to [18].

In each replay step, we sample a batch of experiences from the replay buffer and compute a mini-batch Expected-SARSA update.

Here is the pseudo-code for performing the updates:

Algorithm 2: Deep Expected SARSA with experience replay

$Q_t \leftarrow$ action-value network at t

Initialize $Q^1_{t+1} \leftarrow Q_t$ $N \leftarrow$ Number of replay steps

for $i \leftarrow 1$ **to** N **do**

  $(s, a, r, s', t) \leftarrow$ Sample tuple from experience replay buffer

  if $s \in \mathcal{S}+$ **then**

$$Q_t : Q^{i+1}_{t+1}(s,a) \leftarrow Q^i_{t+1}(s,a) + \alpha \cdot [r + \gamma(\sum_b \pi(b|s')Q_t(s',b))]$$

  else

$$Q_t : Q^{i+1}_{t+1}(s,a) \leftarrow Q^i_{t+1}(s,a) + \alpha \cdot [r + \gamma(\sum_b \pi(b|s')Q_t(s',b)) - Q^i_{t+1}(s,a)]$$

$$Q_{t+1} \leftarrow Q^N_{t+1}$$

2.13. Solar Irradiance and Wind speed simulation

The second component of the methodology proposed is the time series simulation of solar irradiance, air temperature, and wind speed. These variables are needed in order to estimate renewable energy production at any hour of the year. The purpose of including simulation of weather variables in the state space of the RT agent is to provide the agent with information regarding the joint distribution function of solar irradiance and air temperature as well as wind speed as explained in section 2.4. This added information will provide the neural network in the RT agent valuable information regarding the uncertainty and variability of the underlying process of these weather-related variables that directly affect its performance. Instead of using forecasted values, we take advantage of the whole stochastic structure of the time series provided by the copula autoregressive methodology. Usually, forecasting strategies focus on modeling the mean value of the process; moreover, they do not emphasize modeling the underlying variability of the phenomena. furthermore, we also use the copula autoregressive methodology to generate hourly synthetic data of solar irradiance, air temperature, and airspeed for the definition of the *a priori* DA energy commitment parameters for the respective distributions $P^{(k)}_{\psi}$

2.14. COPAR: Copula Auto-regressive simulation

We use the methodology proposed by [19] to generate synthetic data of wind speed time series and the methodology proposed by [20] to generate synthetic data of solar irradiance and air temperature. These methodologies exploit the flexibility of vine copulas for non-linear and asymmetric modeling of serial and between-series dependence. The fundamental pieces to build these autoregressive models are the bivariate copulas, which are distributions on the unit square $[0,1]^2$ such that both marginals are uniform U(0,1). Sklar's theorem [21] explains that for any given continuous variables X and Y with joint distribution $F_{X,Y}(x,y)$ and marginals cumulative distribution functions (CDF) $F_X(x)$ and $F_Y(y)$ respectively, there exists a unique copula function $C_{XY}(\cdot,\cdot)$ that connects $F_{X,Y}(\cdot,\cdot)$ to $F_X(\cdot)$ and $F_Y(\cdot)$ via $F_{X,Y}(x,y) = C_{XY}(F_X(x), F_Y(y))$.

3. Application and case study

We selected a region in Colombia north of the equatorial line [Uribia, Guajira, Colombia] in Latitude (10.76°N) and Longitude (73.00°W) with an approximate elevation of 47 meters as our case study. The data set for this place consist of time series from January 2016 to December 2016. The information was downloaded from Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) data base. For solar irradiance, the data was downloaded from Goddard Earth Sciences Data and Information Services Center (GES DISC) product M2T1NXRAD variable `swgnt` (surface net downward shortwave flux [$\mathrm{W/m^2}$]). Similarly, for air temperature time series, we selected data from (GES DISC) product M2I1NXLFO variable `tlml` (surface air temperature [K]). Finally ,

for air speed time series, we selected data from (GES DISC) product M2T1NXFLX variable `speed` (surface air speed [$\mathrm{m\,s^{-1}}$]).

The virtual power plant configuration consists of an aggregation of two Distributed energy resources (DER) and an Energy Storage system (ESS) with a 150MW/194MWh grid-connected battery. The first DER consists of an hypothetical solar farm with an installed capacity of 28.03 MW (76.800 panels with a nominal power of 365 W) whose characteristics are shown in Table 1.

Array parameters	Value
Number of panels	76,800
Panel type	Flat plate
Dimensions [L×W]	1.7 m × 1.016 m
Nominal (Maximum) Power	365 W
Reference Cell Temperature	25°C
Reference efficiency	21.1%
Array surface irradiation NOCT	800 $\mathrm{W/m^2}$
Temperature coefficient	0.003/°C
Ambient Temperature NOCT ,	20°C
Cell Temperature NOCT ,	44°C

Table 1: Photo-voltaic farm characteristics.

The energy generation of a PV array is dependent on the solar irradiance and the panel efficiency η. This efficiency is a function of the irradiance but also of the cell or cell-array temperature and characteristics of the panel technology.

With the equation 3, the panel efficiency is determined using the cell reference efficiency (at Standard Test Conditions, STC, η_r), STC ambient temperature (T_r) and Nominal Operating Cell Temperature, NOCT, measurements as irradiance ($I_{array,NOCT}$), air temperature ($T_{a,NOCT}$) and cell temperature ($T_{c,NOCT}$).

$$\eta = \eta_r[(1-0.9)\beta\frac{I_{array}}{I_{array,NOCT}}$$

$$(T_{c,NOCT} - T_{a,NOCT}) - \beta(T_a - T_r)] \qquad (6)$$

The electrical energy output Q_e of the array is given by the equation 7. In addition to irradiance and efficiency, we include the number of panels in the solar array N_{panels}, the total surface area of a panel A and the exposition time ($\Delta t = 1$ hour):

$$Q_e = N_{panels}\eta A I_{array}\Delta t \quad \text{[Wh]} \qquad (7)$$

Using this methodology, it is possible to directly generate hourly electricity output data from GHI and air temperature synthetic series. The second DER also consist of an hypothetical wind farm with an installed capacity of 30.00 MW (10 panels with a nominal power of 2.35 MW - 3.00 MM) whose characteristics are detailed in Table 2. The energy generation of the wind

Wind turbine characteristic	Value
Nominal power	2.3 MW / 3.0 MW
Wind class (IEC)	IEC IIA
Cut in wind speed	2.5 m/s
Cut out wind speed	25-27 m/s
normal operation Temperature	-10 °C to +40 °C
Grid feed / control system	ENERCON inverter
Hub height (IEC IIA)	59m / 69m
Rotor Diameter	82m

Table 2: Wind farm characteristics.

farm depends on an specific power curve for the wind turbine selected and the wind speed information. We used the precise information of the power curve [22] to directly generate hourly electricity output data.

4. Results

We analyze the performance of the agents and the main results found after applying the proposed methodology. First, we analyze the convergence measures of the RT MDP agent as well as the DA MDP agent. Figure 3 shows the reward for each one of the 4,000 episodes of the run assessed. The DA MDP agent reached convergence following the convergence measurements presented in Algorithm 1. Due to the restrictions in computing power, we selected 4,000 episodes to be a satisfactory run to pre-train the RT agent.

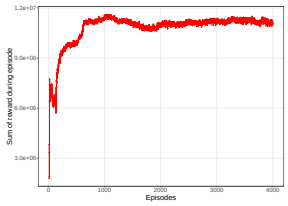
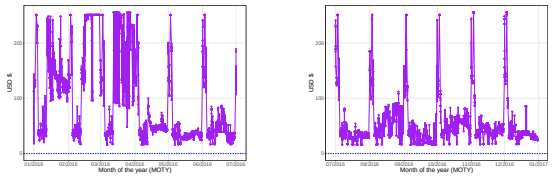


Figure 3: RT MDP Agent sum of rewards per one year episode alongside the agent pre-train run

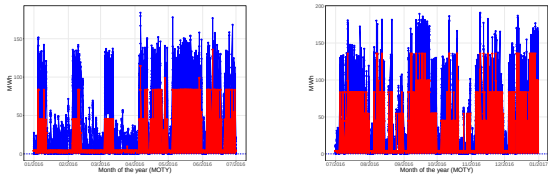
Figure 4 presents the hourly energy price for the case study. The first semester of the year presented one of the most drastic droughts in the history of Colombia. This weather phenomenon caused a dramatic increase in prices towards the first quarter of the year due to the lack of energy production capacity from the hydro-electric plants in the country. Prices eventually normalized once the ENSO season finished. We decided

to select this year (2016) to showcase the capabilities of the methodology to take advantage of the inherent variability of the problem.



(a) First Semester (b) Second Semester

Figure 4: Hourly Electricity closing energy price, Colombia [7]



(a) First Semester (b) Second Semester

Figure 5: Hourly Energy Storage System (ESS) State of Charge (SOC) (Blue) and Energy market purchase (Red)(The reader is referred to the online version for the color palette)

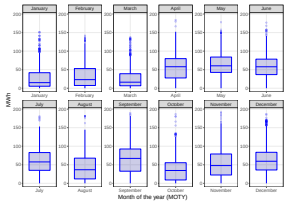


Figure 6: Aggregated monthly ESS state of charge

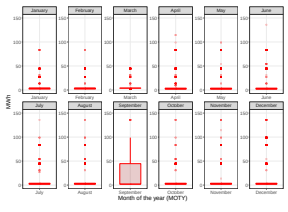


Figure 7: Aggregated monthly RT MDP Market Purchase

Figure 5, 6 and 7 shows the hourly state of charge (SOC) and respective energy market purchases with our proposed methodology. The SOC of the ESS is directly affected by the reduction in energy purchase by the VPP due to the increment in market energy prices experienced in the first quarter of the year as explained before. This price increment creates an opportunity for the agent to takes advantage of the volatility in the market prices to empty its reservoirs and increase its DA energy commitment (Figure 8). Moreover, towards the end of the first semester and during the rest of the year the SOC is capable of responding to the day ahead requirements. In addition, the seasonality trend in Figures 5, 6 and 7 is reflected in the hourly SOC (Figure 5) as well as the energy market purchase in the later months of the first and second semester of the year. This seasonality is mainly generated by the fluctuations in energy production; moreover, this behavior is also affected heavily by the price increments that occur in the first days of every month as shown in Figure 4. Figures 8 and 9 present the hourly parameters (mean and variance) of the DA energy commitment load distributions. Overall, we can see how the agent is adapting to the changing environment without extensive knowledge of the dynamics of the components in the VPP or the mechanics of energy market pricing formation. Figure 11 presents the hourly reward for the proposed methodology. This reward is consistent with the environment state and agent dynamics. The total reward of this exercise for the year 2016 is USD$ 18.5MM. Furthermore, when we analyze both the decisions of the RT MDP agent and the DA MDP agent, both agents make preventive decisions to account for the high variability of the process. This creates a robust DA energy commitment strategy even under extreme conditions that significantly affect the price dynamics of the system.

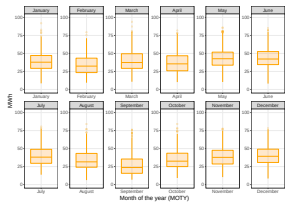


Figure 8: Aggregated monthly day ahead mean from load distribution ψ_i

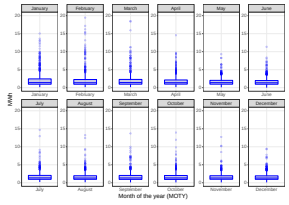
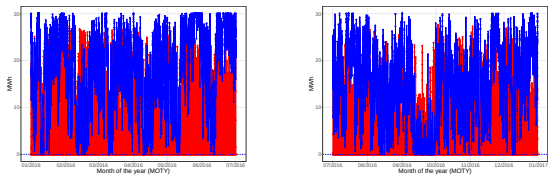
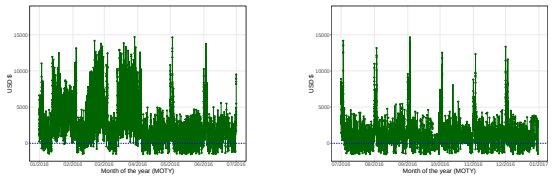


Figure 9: Aggregated monthly day ahead variance from load distribution ψ_i



(a) First Semester (b) Second Semester

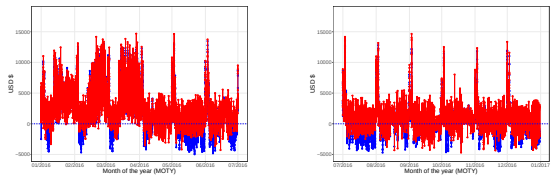
Figure 10: Renewable energy hourly production: Wind farm energy production (red) Solar farm energy production (blue) (The reader is referred to the online version for the color palette)



(a) First Semester (b) Second Semester

Figure 11: Hourly virtual power plant (VPP) Reward

For the sake of comparing and benchmarking the results presented here, we generate a hypothetical agent (naive agent) for the day ahead energy commitment control problem. This hypothetical agent will use the simulations of weather variables to estimate the renewable energy production for the next day as a proxy for the day ahead energy commitment load. This day ahead actions would be supported by real-time actions regarding when and how much energy to purchase from the market. Using the same VPP environment and in order to make this agent realistic, we coded the agent to buy 30% of the installed energy storage capacity whenever the energy price is below a certain threshold. This threshold is estimated as a percentile of the historic energy price. Whenever the price is below the median of the historic price the agent will purchase a specific load to the energy market. Although the decision of this agent is not sophisticated, we want to compare a deterministic approach with a threshold policy with the proposed methodology. Figure 12 presents the hourly reward comparison between the two agents. This comparison reflects on how the naive agent takes advantage of the price increment for the first quarter. However, in a more normal scenario (second semester) the naive agent could not adapt to the changing environment and gets penalized by this.



(a) First Semester (b) Second Semester

Figure 12: Hourly virtual power plant (VPP) Reward comparison. Proposed methodology (red) and Naive agent (blue) (The reader is referred to the online version for the color palette)

5. Discussion and Conclusions

We apply (i) An implementable method for peer-to-peer resource sharing that generates optimal day ahead scheduling policies (ii) A Deep reinforcement learning methodology using the temporal difference method expected SARSA with experience replay for day ahead action validation in real-time. This is assessed integrating two hierarchical Markov decision process using the constrained cross entropy approach (iii) a multivariate copula autoregressive algorithm for simulating solar irradiance, air temperature, and wind speed to solve the optimal control problem on a theoretical day-ahead unit commitment virtual power plant (VPP) environment. The results can provide a better understanding of the VPP day ahead optimal control problem. The copula autoregressive methodology allowed us to generate insightful *a priori* distribution for the DA agent and simulate valuable information of the weather phenomena underlying stochastic process for the RT agent training. The three main advantages of the applied methodology are: (i) The reinforcement learning algorithm is flexible and capable to adapt to ever changing environments with limited previous knowledge of the system and taking advantage of the variability of the phenomena. (ii) The trained agents are able to contribute to the integration of agents in a smart grid. (iii) The use of simulation strategies that focus on understanding the underlying stochastic nature of weather variables rather than just modeling the mean of the process brings insightful information to the decisions makers e.g., the system operator (SO) in the virtual power plant environment.

In our case, the use of an ESS in combination with VRES magnified the benefits of using renewable sources in the hypothetical VPP environment designed. Moreover, the estimation of optimal day-ahead energy commitment distributions could help the system operators to better understand the requirements of the network and work with a better understanding of the underlying stochastic nature of the optimal control problem. In addition, countries that seek to diversify the electricity generation portfolio, such as Colombia, could use the bast amount of natural resources to generate

a more reliable and robust network that is not so dependent on one particular energy source as well as support the decarbonization of the electricity system.

References

[1] P. Corredor, U. Helman, D. Jara, and F. A. Wolak, "Inter-american development bank-idb world bank-bm," tech. rep., Inter-American Development Bank-IDB World Bank-BM, 2020.

[2] A. Lamadrid, T. Mount, W. Jeon, and H. Lu, "Is deferrable demand an effective alternative to upgrading transmission capacity?," *Journal of Energy Engineering*, p. B4014005, 2015/02/05 2014.

[3] R. M. Kretchmar, *A synthesis of reinforcement learning and robust control theory*. Colorado State University Fort Collins, CO, 2000.

[4] A. Tampuu, T. Matiisen, D. Kodelja, I. Kuzovkin, K. Korjus, J. Aru, J. Aru, and R. Vicente, "Multiagent cooperation and competition with deep reinforcement learning," *PloS one*, vol. 12, no. 4, p. e0172395, 2017.

[5] L. Buşoniu, R. Babuška, and B. De Schutter, "Multi-agent reinforcement learning: An overview," in *Innovations in multi-agent systems and applications-1*, pp. 183–221, Springer, 2010.

[6] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," 2019.

[7] XM, "Stock market price and shortage for the colombian market. historical hourly information of close price," tech. rep., XM, 2021.

[8] P. Shaman and R. A. Stine, "The bias of autoregressive coefficient estimators," *Journal of the American Statistical Association*, vol. 83, no. 403, pp. 842–848, 1988.

[9] G. Dalal, E. Gilboa, and S. Mannor, "Hierarchical decision making in electricity grid management," in *International Conference on Machine Learning*, pp. 2197–2206, PMLR, 2016.

[10] S. C. Livingston, E. M. Wolff, and R. M. Murray, "Cross-entropy temporal logic motion planning," in *Proceedings of the 18th International Conference on Hybrid Systems: Computation and Control*, pp. 269–278, 2015.

[11] S. Mannor, R. Y. Rubinstein, and Y. Gat, "The cross entropy method for fast policy search," in *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 512–519, 2003.

[12] P.-T. De Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, "A tutorial on the cross-entropy method," *Annals of operations research*, vol. 134, no. 1, pp. 19–67, 2005.

[13] I. Szita and A. Lörincz, "Learning tetris using the noisy cross-entropy method," *Neural computation*, vol. 18, no. 12, pp. 2936–2941, 2006.

[14] M. Wen and U. Topcu, "Constrained cross-entropy method for safe reinforcement learning," *IEEE Transactions on Automatic Control*, 2020.

[15] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[16] S. Adam, L. Busoniu, and R. Babuska, "Experience replay for real-time reinforcement learning control," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 2, pp. 201–212, 2011.

[17] O. Nachum, M. Norouzi, K. Xu, and D. Schuurmans, "Bridging the gap between value and policy based reinforcement learning," in *Advances in Neural Information Processing Systems*, pp. 2775–2785, 2017.

[18] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *arXiv preprint arXiv:1312.6120*, 2013.

[19] C. Sarmiento, C. Valencia, and R. Akhavan-Tabatabaei, "Copula autoregressive methodology for the simulation of wind speed and direction time series," *Journal of Wind Engineering and Industrial Aerodynamics*, vol. 174, no. 4, pp. 188–199, 2018.

[20] A. F. Ramírez, C. F. Valencia, S. Cabrales, and C. G. Ramírez, "Simulation of photo-voltaic power generation using copula autoregressive models for solar irradiance and air temperature time series," *Renewable Energy*, 2021.

[21] A. Sklar, "Fonctions de répartition à n dimensions et leurs marges," *Publications de l'Institut Statistique de l'Université de Paris*, vol. 8, 1959.

[22] ENERCON, "Enercon e-82 e4 power curve," 2021.