Reinforcement Learning Based Adaptive Power Pinch Analysis for Energy Management of Stand-alone Hybrid Energy Storage Systems Considering Uncertainty

Nyong-Bassey Bassey Etim[a,1], Damian Giaouris[a], Charalampos Patsios[a], Simira Papadopoulou[b,c], Athanasios I. Papadopoulos[b], Sara Walker[a], Spyros Voutetakis[b], Panos Seferlis[d], Shady Gadoue[e]

[a] *School of Engineering, Newcastle University, Newcastle NE1 7RU, United Kingdom*
[b]*Chemical Process and Energy Resources Institute, Centre for Research and Technology Hellas, 57001, Thessaloniki, Greece*
[c] *Department of Automation Engineering ATEI, Thessaloniki, Greece*
[d] *Department of mechanical Engineering, Aristotle University of Thessaloniki, 54124, Thessaloniki, Greece*
[e] *Aston University, School of Engineering and Applied Science, Birmingham, United Kingdom*

ABSTRACT

Hybrid energy storage systems (HESS) involve synergies between multiple energy storage technologies with complementary operating features aimed at enhancing the reliability of intermittent renewable energy sources (RES). Nevertheless, coordinating HESS through optimized energy management strategies (EMS) introduces complexity. The latter has been previously addressed by the authors through a systems-level graphical EMS via Power Pinch Analysis (PoPA). Although of proven efficiency, accounting for uncertainty with PoPA has been an issue, due to the assumption of a perfect day ahead (DA) generation and load profiles forecast. This paper proposes three adaptive PoPA-based EMS, aimed at negating load demand and RES stochastic variability. Each method has its own merits such as; reduced computational complexity and improved accuracy depending on the probability density function of uncertainty. The first and simplest adaptive scheme is based on a receding horizon model predictive control framework. The second employs a Kalman filter, whereas the third is based on a machine learning algorithm. The three methods are assessed on a real isolated HESS microgrid built in Greece. In validating the proposed methods against the DA PoPA, the proposed methods all performed better with regards to violation of the energy storage operating constraints and plummeting carbon emission footprint.

Keywords: Hybrid Energy Storage Systems; Energy Management Strategies; Model Predictive Control, Kalman Filter; Reinforcement Learning

---

[1] Corresponding Author at: School of Engineering, Newcastle University, Newcastle NE1 7RU, United Kingdom.
E-mail address: B.E.Nyong-Bassey1@ncl.ac.uk

The short version of this paper was presented at ISCAS2018, May 27-30, Florence, Italy. This paper is a substantial extension of the ISCAS2018 Conference paper.

Nomenclature

$AEEND$	Available excess energy for the next day	Δk	Time interval
BAT	Battery	δ	The proportion of flow j
C_l	The capacity of accumulator l	$\eta_{CV}, \eta_{PV}, \eta_{FC}, \eta_{EL}$	DC converter, PV panel, fuel cell, electrolyser efficiency factors
DSL	Diesel generator	$\varepsilon_i(k)$	Binary variable for the state of the ith dispatchable unit
EL	Electrolyser	$\rho_i^{i_c}$	The binary variable related to the temporal conditions of the accumulator
FC	Fuel cell		
		Subscripts/superscripts	
HT	Hydrogen Tank	$SOAcc$	Accumulator or energy storage
G	A fixed reward	Avl	Availability of resources
$\mathcal{J}$	Identity matrix $\in \Re^{nxn}$	Gen	Override logic for PoPA energy dispatchable units *FC* and *EL*
LD	Load	Req	Demand for resources
MAE	Minimum absorbed energy	k	Time step
$MOES$	Minimum outsourced energy supply	i	Index of Converter
s^-	Previous state before a transition by the agent	l	Accumulator
$SOAcc_l^n$	State of accumulator l	max	maximum
S_{Lo}^l	Lower pinch limit or utility	min	minimum
S_{Up}^l	Upper pinch limit or utility	m, n	Model and the plant respectively
POW	Power flow	i_c	A set of controllable energy converter elements for PoPA targeting
$PGCC$	Power grand composite curve	$\rightarrow$	The arrow head indicates the direction of flow of energy/material from source to sink
$\mathcal{R}$	Zero mean Gaussian noise $\in \Re^{nxn}$		
$\mathcal{U}$	Input $\in R^{mx1}$		
$W1, W2$	Penalty weights which control the propagation of the negative reward exerted on the agent.		
WT	Water tank		

1. Introduction

Growing concerns over the impact of greenhouse gas emission on the environment has led to policy initiatives to advance the proliferation of renewable energy sources (RES) (such as wind turbines and solar panels), for distributed generation (DG). Furthermore, in remote areas without access to an electrical grid, RES are a favourable electrification alternative when compared to the cost of deploying high-voltage transmission lines and associated power losses [1-3]. The use of RES (particularly in a standalone microgrid (MG)) can reduce the reliance on backup diesel generators (DSL) which have a high carbon emission impact on the environment [4, 5]. Nevertheless, due to weather stochasticity, some RES can have predictable but variable power output and so, incorporating energy storage technology with RES can mitigate this variability. Multiple energy storage technologies (e.g. battery and hydrogen) with complementary properties (such as life cycle, seasonality, power and energy density etc.) are often combined to further mitigate the RES variability. This is the concept of hybrid energy storage systems (HESS) as shown in Figure 1 [6, 7]. This system was designed and built in Xanthi, Greece in collaboration with CERTH and SUNLIGHT [8] and it is been used here as a case study. The mathematical model of each asset has been previously validated [9] by the authors and real load/weather profiles have been used.

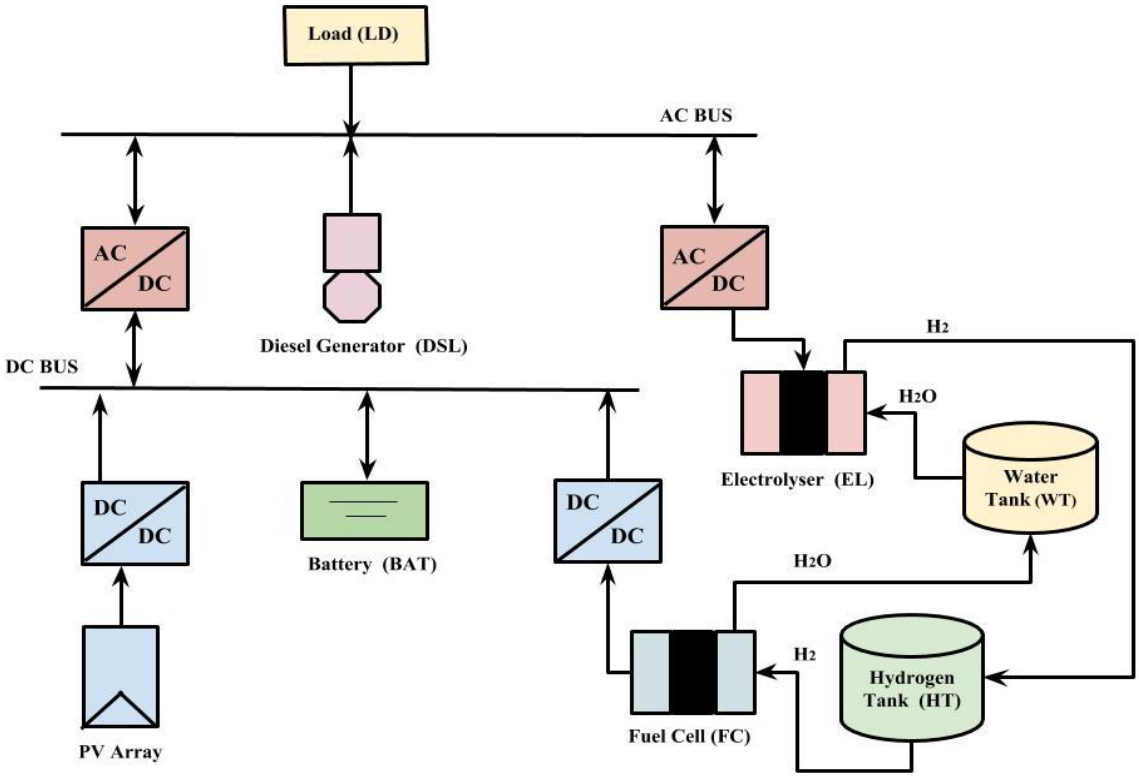


Fig. 1. Schematics of the experimental Islanded HESS [7] used as a case study

In such systems, when supply exceeds demand and a local battery is completely charged, the energy from the RES can, for example, be converted to hydrogen (H_2) by an electrolyser (EL) for long term storage (as opposed to the battery that can be seen as short-term storage option). Then, the hydrogen can be used when demand exceeds

supply, by means of a fuel cell (FC) [7, 10]. The HESS thereby can reduce the dumped load in times of excess supply, and further reduce the need for backup DSL in times of excess demand [11]. A newer innovative hydrogen production approach, which relies on internal rather than external reforming of fuel mixtures into mass production of electric and thermal energy carriers, with high efficiency, based on the use of Solid Oxide Fuel Cells (SOFCs) have recently been investigated. In [12], an intermediate temperature solid oxide electrolyser stack is fed with carbon dioxide (CO_2)-steam mixture at the anode. Here the fuel mixture is reformed into CO - H_2 mixture while at the cathode, oxygen fed into the system is converted into ions. The oxygen ions generate current as they pass through the electrolyte towards the anode where they combine with the CO - H_2 mixture to produce CO_2 and water. The work evaluated the thermal equilibrium current at the highest stack operating temperature for hydrogen production. Furthermore, authors [13] investigated the use of low weight as well as low cost high temperature steam electrolysis (HSTE) stack for durability and performance to highlight current density and steam conversion ratio at the temperature of 800^{o}C. The authors were optimistic about the future of HSTE technology concerning performance, durability, thermal cyclability, as well as low cost. Another innovative method for hydrogen production is the anion exchange membrane (AEM) FC which is attractive due to its outstanding fast electrochemical kinetics, low dependence on non-precious catalyst and water removal mechanisms [14]. The concept of anion exchange membrane (AEM) FC whereby, negatively charged oxygen ions travel from the cathode (negative side of the FC) to the anode (positive side of the FC) instead of positively charged hydrogen ions traversing from the anode to the cathode, as is the case in all other types of FCs is increasingly being applied as an innovative method for hydrogen production. The movement of the anions and reactions at the anode produces electricity and water as by-products and can be recycled for anion and hydrogen production. In this system creating fast electrochemical kinetics and water management are essential for sustainable operation. In [15] an analytic model for alkaline anion exchange membrane FC is proposed. The authors in their investigation, illustrated more anode humidification improved performance. However, at higher ranges of humidification levels this improvement became less significant. Nevertheless, a systems-level analysis approach which can be generalised in principle to a broad range of energy systems has been implemented in this work, hence, the impact on the HESS as a result of integrating these newer H_2 technological innovations which were highlighted will be an interesting subject for future investigation.

Despite the advantages offered by a HESS, the heterogeneity of the components/devices introduces complexity due to the need to account for different forms/characteristics of energy flows between multiple assets and for numerous decision parameters in energy management strategies (EMSs) used for HESS control. To address such complexity, several studies have proposed the use of if-then-else rules, artificial intelligence (AI) (such as fuzzy logic controllers, neural networks, and genetic algorithms), linear and dynamic programming and advanced control techniques to realise EMSs for HESS [16-18]. Development of EMSs using if-then-else rules in the form of hierarchical diagrams is widely used in published literature due to its computational efficiency [16].

In [19] a rule-based EMS was proposed for domestic microgrid. Specifically, the predicted load and PV generation power, utility cost are utilised in conjunction with the batteries state of charge as the input of the rule based algorithm at each interval. Thus, the rules are such that the load requirement at each time interval is compared with the PV power and which only fulfils the load power requirement, whenever the output power of the PV is greater and given the battery level, any excess is either used for charging operation or arbitrage. Then again, if the load demand exceeds the PV power, given the cost of the battery pack and utility determines and battery level is

used by the EMS to decide how to cover the deficit. Thus, the rule based EMS had accurate result and faster processing time in comparison with an optimisation based EMS. However, this approach is largely heuristic and limited to very few potential options, omitting numerous alternatives which may improve the HESS performance, as illustrated in [7]. In addition, fuzzy logic controller which is classically rule-based has enhanced adaptation and robustness in contrast to a conventional rule base controller as depicted in the case of energy management (EM) of islanded MG in [20].

In [21] fuzzy logic control strategy comprising self-organising fuzzy logic and fuzzy dynamic decision making was used to estimate the required output power of a FC based on the driving load requirement and state of charge of a BAT in an electric vehicle (EV). Furthermore, in MATLAB®/ Simulink®/State-flow simulation environment, the proposed strategy was shown to improve the efficiency of the EV. In [22], the merits underling the integration of hybrid energy systems, specifically; a FC, BAT and supercapacitor in an EV are first analysed. Thereafter, an active power flow control technic is proposed based on optimal control theory with the objective of optimising BAT life and total energy cost while meeting vehicle loads demand requirements based on the minimisation of a square error cost function between the desired and actual parameters. The proposed method was validated against existing control technics had better performance in driving cycles while operating the assets in a suitable manner. In [23] an energy calculation tool is proposed and implemented in the MATLAB® and Simulink® environment for hybrid polymer electrolyte FC based on a generic users predefined route. The calculator tool accounted for electric energy recoverable downhill and in the course of deceleration period. In [24] an optimal control strategy based on a two dimensional Pontryagin's minimum principle, was proposed for EM of a batteries and super-capacitor in a plug-in hybrid electric vehicle. The optimisation approach led to improved battery degradation and a 21.7% reduction in total economic; fuel, electricity outsourcing and maintenance cost. In [25] a dynamic EMS was proposed in response to deviation in dc-link voltage ensuing from dynamic load and RES uncertainty in a grid connected HESS microgrid which comprised a battery bank and ultra-capacitor. In [26] a piecewise robust optimisation EMS was proposed for combined cooling heating and power microgrid with the objective of minimising total cost under the worst case scenario to carter for power uncertainty. In [27] a dual stage robust MPC optimisation is proposed, in order to reduce the impact of load demand and RES uncertainty in an islanded MG. In the first stage, operational cost under a joint worst case scenario is minimised. Thereafter, with the observation of actual data, minimisation of the adjustment cost is performed with an economic dispatch model. However, robust optimisation method is considered as a pessimistic approach and can result in over budgeting in real world application [28]. More so, stochastic and chance constrained based optimisation which have been applied in **[**29-32] and [33 -35] respectively for Energy management of MGs are not only computationally cumbersome and but also intractable. Hence, the use of approximate solutions which largely depend on the accuracy of probabilistic distribution or explicit modelling of the underlying uncertainty parameters, which is practically limiting in real-world applications as the distribution might be unavailable [26, 34]. Furthermore, in [36] MPC strategy with corrective feedback was proposed for energy management of a domestic microgrid was shown to achieve better energy savings than the standard rule based control strategy. In [37] MPC combined with adaptive-Markov chain prediction was proposed for energy management of a dual hybrid EV. The MPC based method achieved better fuel economy over a rule base strategy. In [38] real-time EM optimal control algorithm for a dual mode split HEV formulated as a multivariate quadratic optimisation problem solved offline to obtain control laws which was thereafter applied in real time in a traditional MPC manner. The proposed strategy had

reduced computational cost and fuel economy of 97.46% and 23.3% respectively compared to the traditional MPC.

On the other hand, AI or mathematical programming approaches are able to investigate a very large number of options and to identify optimum solutions. However, they may suffer from increased computational demands due to combinatorial complexity or non-linear system models, which makes them inefficient for on-line decision making [39, 40]. Furthermore, they only provide one final solution which hinders the opportunity to derive insights from intermediate solutions and analyse the HESS operation. To address such shortcomings, the Power Pinch Analysis (PoPA) [41, 42] was proposed both as an effective means of graphical EMS analysis and a tool which may enhance the computational efficiency of mathematical optimization approaches. PoPA is a process integration technique, inspired from the original Pinch Analysis for heat exchange networks [43] and evolved to sophisticated tools [44] that allow the analysis of complex energy systems based on the identification of insights pointing toward promising design and operating decisions [45]. The PoPA, used as a graphical and/or numerical tool, aids in the identification of deficit or surplus targets for energy recovery by the use of dispatchable resources to satisfy a conservative minimum energy target. It considers power demand and supply requirements with respect to time in the form of the Power grand composite curves (PGCC) to identify inflection points (called pinches) where power demand must be satisfied by external, non-renewable energy sources or excess power availability will be dumped, unless exploited internally. The identification of pinch points allows the development of EMS which support efficient internal energy recovery so that the use of non-renewable energy or the dumping of renewable energy can be avoided [5]. The PoPA, which has mostly been used for optimal sizing, planning of energy supply and demand management in hybrid energy systems, has recently grown in use compared with mathematical programming techniques [46]. Some of the promising aspects of PoPA are reduced computational effort, analytical insights derived through a graphical interface tool, as well as the systematic consideration of the assets' interdependence and intrinsic complexity [5].

1.1. Applications of PoPA for Electric Power systems sizing and design

Several researchers have considered PoPA for electric power systems sizing and design. In [41, 46] the grand composite curve was realised by integrating the energy demand and supply over time, and then it was used to optimally size an isolated power generation system. Additionally, in [47] the PoPA was utilised as a combination of both the graphical analysis and numerical approach with the aid of the power cascade analysis and storage cascade table for optimal sizing of the hybrid power system. The extended Power Pinch analysis (EPoPA) in [48] was proposed as an enhancement to the PoPA in order to optimally design renewable energy systems integrated with battery-hydrogen assets as well as a DSL. The EPoPA was used graphically and algebraically to determine the required external electricity to be outsourced, the wasted energy which could not be stored in the battery (BAT), but can perhaps be stored in form of hydrogen in a normal operational year. Thereafter, the sizes of the hydrogen tank (HT) and DSL were determined by minimising the total annualised cost. These studies on PoPA for sizing MG assets with the exclusion of [46] in which chance constrained programming was used to achieve technical and economic feasibility, were realised without recourse to uncertainty.

1.2. Applications of PoPA for energy management

Apart from the use of PoPA in electric power systems sizing and design, it has also been used, by the authors, as an EM tool, as first reported in [5, 7, 49]. More specifically, in [7] the power grand composite curve (PGCC) was realised within a model predictive control (MPC) framework for the first time with a day ahead (DA) forecast to infer and effect (EM) decisions in a HESS stand-alone MG. By shaping the PGCC, a series of optimal control decisions for the activation and duration of the HESS operation were determined. Therefore, the EMS was contingent on the identification of the energy recovery targets within the prediction horizon. The effectiveness of this approach was limited by the assumption of a perfect DA weather and load forecast.

1.3. Generic approaches to uncertainty

The pinch analysis despite being a well-established process integration recovery and conservation technique for assets such as waste management, water, heat, and carbon emission requires consideration and expansion in power systems application [42]. Also, as highlighted, most literature on PoPA have not dealt with uncertainty, as these studies have mostly relied on the assumption of perfect (or ideal) weather forecast and load profile with the exception of [46] where uncertainty was considered in the sizing of a MG asset. Consequently, the significant impact of uncertainty, imposes the need to integrate PoPA tools with a complementary technique, especially when consistency is so desired. The techniques which account for uncertainty in EM can fundamentally be classed as either predictive or reactive approach [50]. These predictive or reactive approaches may perhaps be considered in PoPA application, whereby, the scheduling of dispatchable units are realised with or without prior consideration for the impact of an impending uncertainty respectively. The reactive approach uses the latest state feedback for re-computation, upon model mismatch due to uncertainty, which may be expensive when seeking an optimum solution in the event of frequent perturbation. The predictive technique may employ stochastic programming, fuzzy programming, robust optimisation, machine learning techniques, in order to infer the optimal control action that negates the effect of uncertainty [51-53]. Furthermore, the linear Kalman filter, first presented by Kalman in 1960 for solving the Wiener problem has since been applied extensively in areas of control system, short-term prediction, navigation tracking and for systems state estimation associated with uncertainty [54]. In [55] the ensemble Kalman filter was combined with a multiple regression model to enhance forecasting accuracy of electricity load. Similarly, in [56] the Kalman filter was used recursively to estimate short-term hourly load demand forecast parameters based on the historical load and weather data and the current measurements of the time-varying parameters. Moving away from the well-known prediction methods, the work of [57] on temporal difference (TD) learning, a model-free reinforcement learning (RL) algorithm, introduced a prediction method which relies on the experience of successive predictions to infer the behaviour of an unknown system. This was a paradigm shift to the conventional approach which depended only on the difference between the actual and predicted outcome. Hence, RL is a machine learning technique, suitable for solving a Markov decision process (MDP) which involves sequential optimal decision making under uncertainty. Thus, many researchers have sought to deploy several machine learning algorithms in an MDP. In [58], machine learning algorithms such as policy iteration and value iteration Dynamic programming, and RL techniques such as the least squares policy iteration, Q-Learning, and SARSA were reviewed for MDPs. Specifically of interest, is the Q-learning, a class of model-

free RL, a similar algorithm to Sutton's (1988) TD learning [56], first introduced by Watkins in 1989, which proffers an intelligent agent with the learning ability to act optimally in a MDP based on experience [59]. In Q-learning, an agent seeks to maximise the sum of expected reward by acting optimally with respect to any given circumstance (referred to as a state). Typically, an agent will evaluate a state, and will then undertake an action either in an exploitative or exploratory manner thereafter and finally will receive an instant reward, while transitioning to a new state. Q-learning has tremendous success in robotics, especially in mobile robot navigation and obstacle avoidance [60, 61]. In [62] the Dyna AI architecture was proposed to integrate both learning, and experience, based on online planning, as well as reactive execution in a stochastic environment.

Furthermore, in [63] a comparative study of MPC and Monte Carlo RL on a non-linear deterministic system with known uncertainty dynamics was undertaken. The two methods were compared with respect to three cases; linear, uncertain and/or stochastic. The author noted that for linear systems, the MPC performs better than the RL as it converges to a convex solution, while the RL suffers from suboptimality while tracking an available trajectory. In an uncertain system, the RL is capable of inferring optimal policy from real life or available trajectories despite poor information which may abound in such targets, while the MPC may act sub-optimally due to open loop. Furthermore, for a stochastic system, the open loop policy of the MPC is sub-optimal, a problem due to lack of feedback which RL does not suffer from. However, in both cases, robustness in MPC may be improved by the addition of a receding horizon. More recently, [64] harnessed the merits of the MPC and RL control strategies to form an adaptive controller for a heat pump thermostat based on the suggestion of [63]. The adaptive controller maximised energy savings while tracking a varying temperature set-point for thermal comfort, more effectively than the MPC or RL alone. The strategy employed MPC for planning the optimal control action corresponding to each state at initialization with the assumption that thereafter, RL is used to update the dynamic model online.

The application of RL based energy management for HESS has mostly been considered in literature with respect to hybrid Electric vehicle while only a few have considered microgrid systems. In [65] energy management based on a 2 steps-ahead RL framework was proposed for a grid connected microgrid which comprised consumers load, ES, wind turbine. The RL is formulated as a multi-criteria decision making tool, aided by a 2 steps-ahead prediction of available wind power via a Markov chain model. This approach allowed the learning agent to optimally utilise the WT, independently of the grid to charge the ES. On the other hand, it maximised the use of the ES during peak demands. Hence, enabling an intelligent consumer to learn a stochastic scenarios while incorporating experience based optimal actions. In [66] deep RL EMS which uses a convolution neural net to extract relevant time series information, from a large continuous non-handcrafted feature space is proposed to address stochastic electricity production in a residential MG. The neural net is validated periodically during training on historical features of observation to reduce over fitting and positive bias. The levelized energy cost economic criteria with respect to maximizing operation revenue is used to evaluate the performance of the algorithm. In [67] the authors propose an EMS which applies a decentralised cooperative multi-agents enabled Fuzzy Q-learning to a standalone MG. The formulation of the continuous input states entails the use of five membership functions and the action space comprising a fuzzy set pertaining to each MG asset and rules base in conjunction with a reward formulation, shapes the agent's continuous action policy. In [68] the authors proposed a real-time EM algorithm to optimise performance and energy efficiency with power split control for a hybrid (battery and ultra-capacitor) tracked vehicle for various road driving conditions. A speedy Q-Learning algorithm

is used to accelerate the convergence of a multiple transition probability matrix which is also updated whenever the error norm exceeds a set criteria. The proposed method which was compared to a stochastic dynamic programming approach and a conventional RL using two driving cycles had an improved fuel economy. In our work we have excluded the use of a Markov chain to model a stochastic transition probability matrix (TPM) of the MDP, as this not mandatory in the development a RL framework [69]. Though in [70] and [68] Markov chain is used to model a stochastic TPM which is updated periodically when a specific criterion is exceeded by the magnitude of an induced matrix norm and kull-back divergence respectively. This is in contrast to an earlier proposed method in [71] where the authors for the first time applied reinforcement learning technique (specifically TD(λ)) to minimise the fuel consumption of a hybrid electric vehicle without the need for prior knowledge or stochastic information of the driving cycle, and uses only a partial hybrid electric vehicle model. Nevertheless, our proposed RL formulation requires only the (corrected) adaptive Pinch analysis target, strictly for evaluating the environment state and scalar reward which the dyna-Q learning agent receives after taking an action in a given state. Furthermore, the step wise non-linear optimisation used to derive the optimal control strategy in [70] and [68] and a backward-looking optimisation in [71] is replaced with a heuristic graphical based adaptive power pinch analysis MPC framework, which we have proposed in our work. Thus, eliminating the computational cost associated with building a TPM offline, as well as solving a complex non-convex optimisation EMS for HESS (particularly with heterogeneous energy and flow mix as in our case, where we have to deal with the intrinsic interaction of power, hydrogen, and water flow between subsystems). Furthermore, we have omitted detailed operational considerations with regards to losses associated with device level operation, since the considered EM approach is at the systems level.

Nevertheless, evaluation and formulation of the scalar reward in aforementioned RL papers excluding [70] which applies a backward-looking optimisation, have mostly been implemented subjectively and without recourse to a systematic approach which determines the ideal optimal action strategy as in the use of a corrected adaptive PoPA. Hence, these rewards are based on a local maximisation which increases the operational cost and incurred excess energy losses in contrast with a global maximum insight which the corrected adaptive PoPA offers.

1.4. Main Contributions and Novelties

It is clear that PoPA has rarely addressed the issue of uncertainty and only in a case of HESS sizing, while the PoPA approach has significant advantages (described above) in cases of adaptive EM. To this end, such advantages have been previously exploited by the authors within an MPC framework, however under limiting assumptions of perfect weather and load forecasting. The focus of this work is therefore on addressing the issue of RES/load forecast error which is bound to occur in a realistic scenario, in the context of the PoPA approach. Three novel adaptive PoPA schemes are proposed based on an EMS algorithm for an islanded HESS aimed at significantly reducing the effect of forecast error while shaping the PGCC. It has to be noted here that the islanded HESS that is being used here as a case study, has been designed and built by the authors at CERTH in collaboration with SUNLIGHT [8], and the mathematical models of the assets have been previously experimentally validated [9].

More specifically, the main contributions of this work are as follows:

I. The DA PoPA in [49] for EM of HESS has been adapted for the first time, to realise an 'Adaptive PoPA' [72], by re-shaping the PGCC in a multi-step, look ahead, receding horizon MPC framework as shown in Figure 2. This method offers a simple but effective means to counter the effects of forecast error.

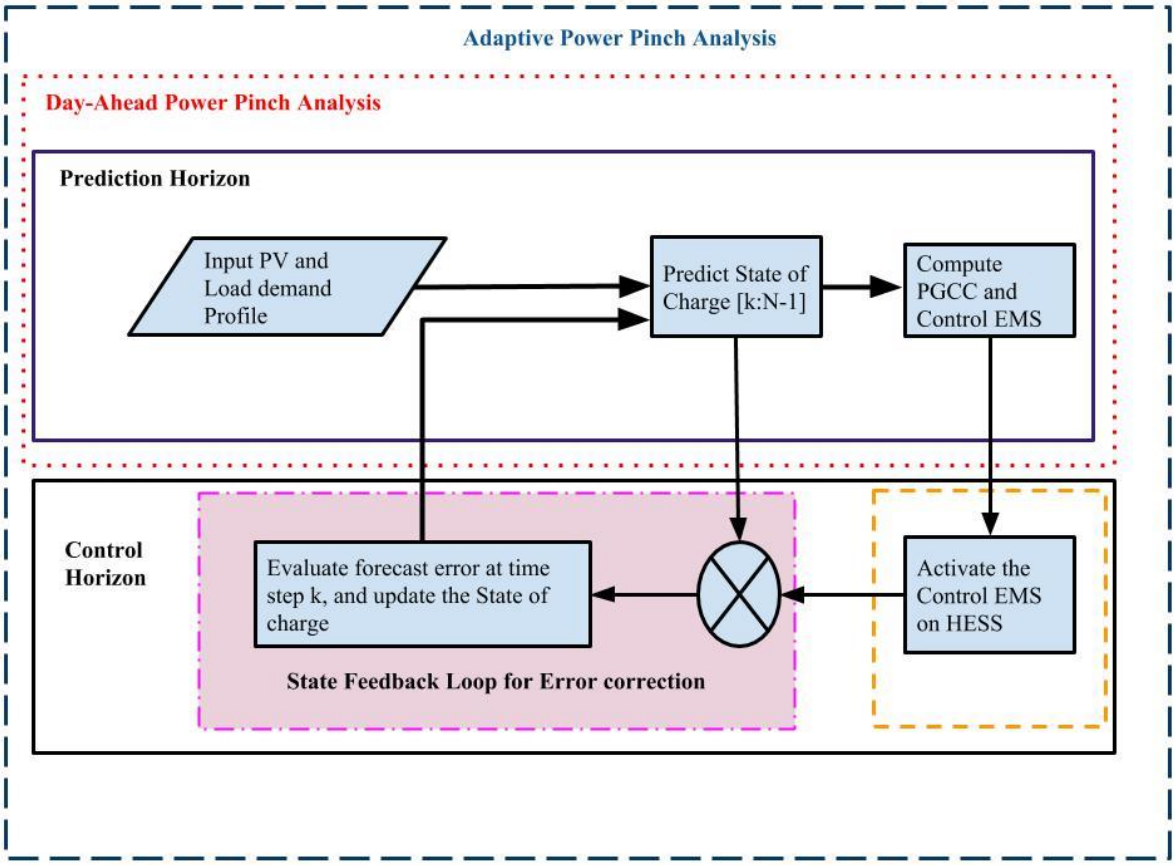


Fig. 2. Schematics of the Adaptive Power Pinch Analysis EMS for HESS [40]

II. A Kalman filter for the first time, has been used in conjunction with the aforementioned Adaptive PoPA [72], to predict the State of Charge of the battery ($SOAcc_{BAT}^{m}$) based on the likelihood estimation of uncertainty. The algorithm is more sophisticated than the Adaptive PoPA but nevertheless computationally efficient and offers a preventive measure as an improvement. Furthermore, the occurrence of the forecast error is not dependent on the corrective action, as in case (I), which may improve the algorithmic performance.

III. A RL-based adaptive PoPA (RL+Adaptive) method has been proposed for the first time, in the context of the *dyna* Q-learning algorithm. The *dyna* Q-learning algorithm entails learning a policy by means of rewarding an agent based on the next state of the system after inferring a control action given the current state of the system. Thus, the agent learns an EMS by solving for the optimal action policy. Additionally, with the action policy, the agent decides the de/activation of the dispatchable units in accordance with a corrected PGCC shaped with the Adaptive PoPA. This approach does not assume that the underlying uncertainty is normally distributed in the

procedure that minimizes the mean squared error in the estimated state-of-charge, as in case (II). This may improve the algorithmic performance, hence it is worth investigating.

The three approaches are analysed in this paper. Furthermore, a sensitivity analysis with hydrogen uncertainty is used to evaluate the proposed methods against the DA PoPA. The rest of the paper is structured as follows: Section 2 briefly describes the Power Pinch concept. Section 3 presents the formalisation of the receding adaptive MPC-PoPA concept. In section 4 and 5, the proposed Kalman filter state estimator approach with Adaptive PoPA and the RL Adaptive PoPA algorithms are presented, respectively. The results are presented in Section 6, and Section 7 provides a conclusion.

2. Power Pinch Analysis for Energy Management of Hybrid Energy Storage Systems

2.1 Generic description

In order to understand how Pinch Analysis can be used to determine an EMS in a HESS (as shown in Figure 1), infer a generic islanded energy system with multiple energy carriers (like electrical and hydrogen), multiple storage assets (like a BAT and a HT), generation assets (like photovoltaic panels (PV)), controllable assets that can transform an energy from one carrier to another (like a FC and an EL) and a load (possibly for each energy carrier). Also, for each storage component we set up operating limits that should not be violated, say S_{LO} and S_{UP} which is the minimum and maximum allowed stored energy/material respectively.

The first step to apply the PoPA concept is to define the Power Grand Composite Curve (PGCC) for each energy carrier, which is the integration of all uncontrolled energy demands and generation in the system for that carrier for each instance. When the system is at a specific instant *k*, we predict the PGCC as shown in Figure (2a) by assuming that the controllable assets are not activated and we check if the predicted PGCC violates any of the aforementioned limits. The predictive horizon is based on an hourly interval which spans for 24h $\in [k: N]$, where k is the i^{th} hour in a day and N indicates the end of the day (or 24$^{\text{th}}$ h). The hourly interval Δk is expressed as the difference between two successive time steps; $\Delta k = [(k+1) - k]$ where, k and $k+1$ are the current and next time step respectively. The interval between the current time step k and the end of the horizon N is given as $(N - k)/\Delta k$, and the entire horizon would have 23 intervals, if k is the first hour, 01: 00h and $N = (k + 23)$ is the 24: 00h of the day. If the PGCC violates a limit at a specific instant, then at an appropriate instant before the violation occurs, a suitable controlled asset will be activated in a control horizon of interval 24h $\in [k: N]$ with equivalent time duration as in the predictive horizon in order to provide/remove the necessary energy/material so that the system limits are not exceeded. In order to better describe the aforementioned concepts, a specific motivating case will be presented in the next subsection.

2.2 Motivating case

In the HESS as shown in Figure 1, let the stored electrical energy (i.e. state of charge, $SOAcc$) be the quantity that we wish to control within specific operating limits. Therefore, an EMS is derived in prediction horizon using a DA strategy and implemented on the HESS in a control horizon. In the prediction horizon, $SOAcc$ is plotted

(dotted black line in Figure 3a) at an hourly time step k, for a daily (24 h) span as defined in section 2.1. The PoPA enables the identification of deficit and excess energy targets, which must be successively met, in order to prevent the $SOAcc$ in the control horizon from falling below the lower pinch utility (or limit) S_{Lo} (say 30%) and/or rising above the upper pinch utility S_{Up} (say 90%).

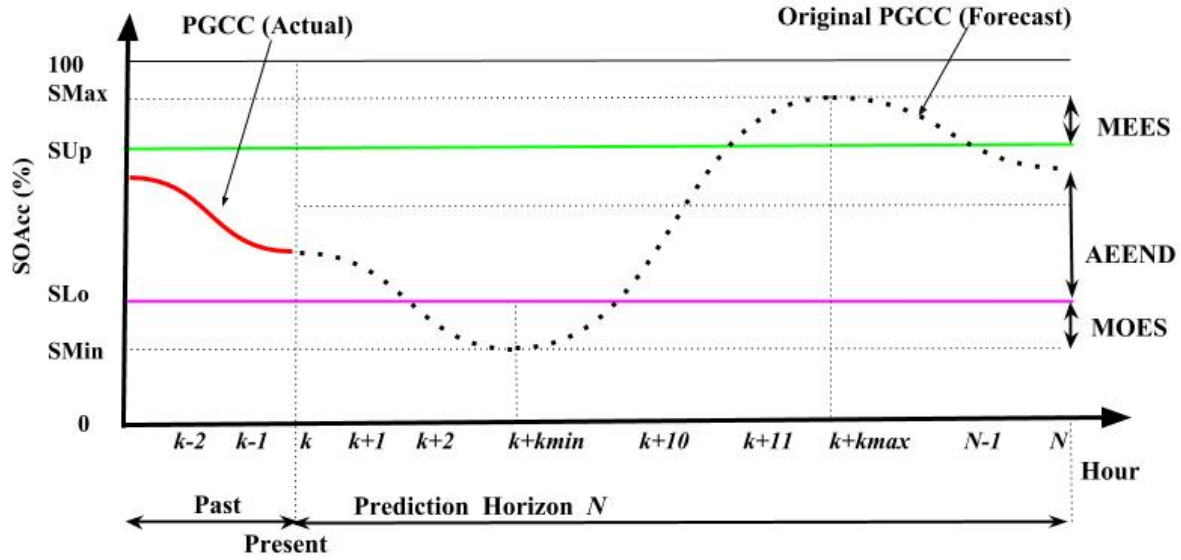
At first, the control strategy aims to determine the deficit energy target at the minimum $SOAcc$, denoted as $Smin$. In this case study, the deficit results from the absence of sufficient energy supply by the PV. The deficit energy target is then the amount of energy needed to ensure $SOAcc$ avoids the violation of the S_{Lo} limit at time $k + kmin$. The PGCC determines the minimum amount of outsourced electricity supply (MOES) required in order to violate S_{LO}. A dispatchable asset, (such as a FC) indicated by a red arrow pointing upward at time k shown in Figure 3b, supplies the energy needed to shift the PGCC above S_{Lo}.

Secondly, the control strategy aims to determine the excess energy target at the maximum $SOAcc$, denoted as $SMax$. The excess energy target is then the amount of energy that needs to be dumped in order to avoid the violation of the S_{UP} limit at time $k + kmax$. This is denoted as the minimum excess energy for storage (MEES). Thus, the MEES is recovered for storage by a dispatchable asset (such as an electrolyser (EL)) denoted by the red arrow pointing downwards shown in Figure 3b.

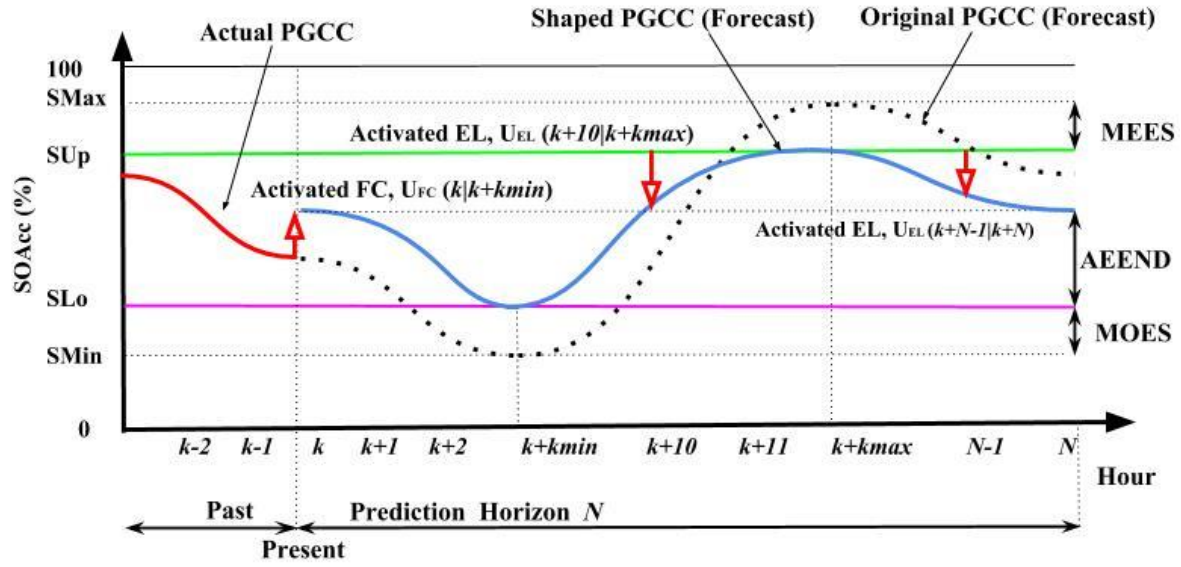
Thirdly, to preserve the duty cycle of the energy storage, the available energy for the next day (AEEND) i.e. $SOAcc$ at time step N has to be matched to the $SOAcc$ at time step k, by activating dispatchable assets (either the FC or EL) at time step $N - 1$.

Consequently, by shifting the entire PGCC up or down (black dot-dashed line in Figure 3b), there are instances where the PGCC reaches (but no longer exceeds) the S_{LO} or S_{UP} at times $k + kmin$ and $k + kmax$, which is termed the Pinch point. Therefore, the shifted PGCC which resolves the PoPA EMS is responsible for the instant and duration for which the energy targeting resources are activated/deactivated in the control horizon [5, 7, 49, 73].

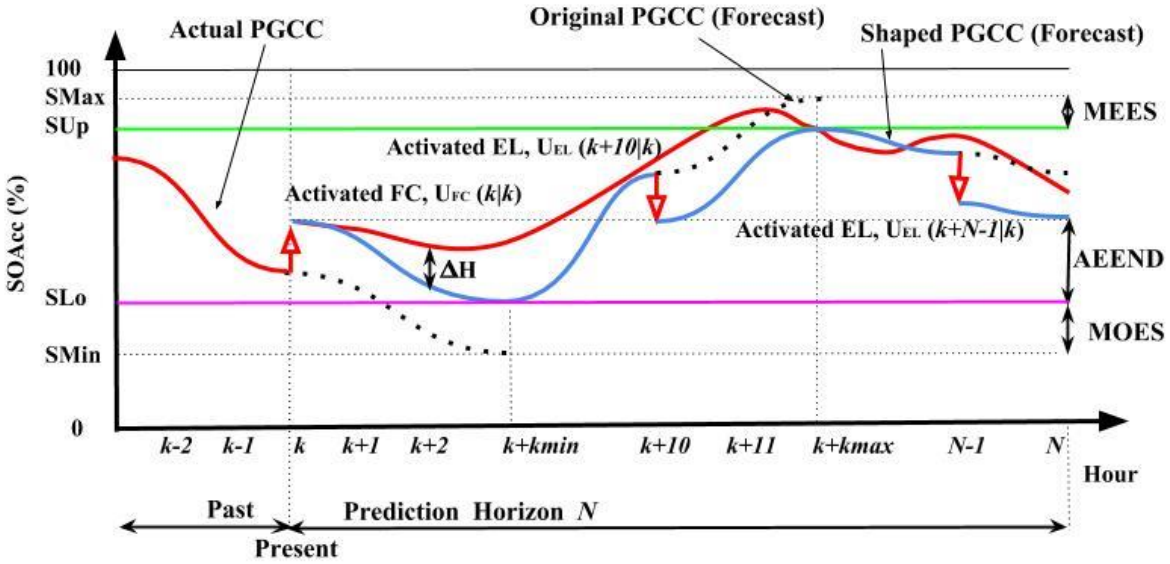
However, effectively realising the optimal PoPA EMS via DA operation requires an accurate load and weather forecast model for an ideal PGCC plot, which is impractical due to uncertainty for most real applications. The effect of uncertainty, ΔH due to RES variability and stochasticity of electricity demand, causes a mismatch between the actual (red line) and predicted (blue line) $SOAcc$ as illustrated in Figure 3c and consequent violation of S_{UP} and the duty cycle constraint. Therefore, the utilisation of a feedback loop is crucial to improve the excess energy recovery and reliability indices. It can also reduce the need for (potentially higher carbon emission) energy imports to the system.


PGCC (Actual)
Original PGCC (Forecast)
100
SMax
SUp
SLo
SMin
0
SOAcc (%)
MEES
AEEND
MOES
k-2
k-1
k
k+1
k+2
k+kmin
k+10
k+11
k+kmax
N-1
N
Hour
Past
Present
Prediction Horizon N

(a)


Actual PGCC
Shaped PGCC (Forecast)
Original PGCC (Forecast)
100
SMax
SUp
SLo
SMin
0
SOAcc (%)
Activated EL, UEL (k+10|k+kmax)
Activated FC, UFC (k|k+kmin)
Activated EL, UEL (k+N-1|k+N)
MEES
AEEND
MOES
k-2
k-1
k
k+1
k+2
k+kmin
k+10
k+11
k+kmax
N-1
N
Hour
Past
Present
Prediction Horizon N

(b)



(c)

Fig. 3. (a) Original PGCC; (b) Shaped PGCC and (c) the effects of uncertainty with the DA-PoPA

3. Adaptive Power Pinch Analysis

The effects of uncertainty on renewable energy sources and electricity demand with respect to the DA-PoPA operation have been highlighted in section 2. Thus, in this section we adapt the DA-PoPA, to create an Adaptive PoPA which uses a receding horizon MPC approach. In a prediction horizon spanning 24 h with hourly interval Δk and time step k, as defined in section 2, the dispatchable control variable $U_c(k)$ is determined based on the PoPA targets. Accordingly, $U_c(k)$ determined in the prediction horizon is activated in control horizon at each time interval k. Furthermore, the $SOAcc$ as a function of the minimum energy recovery is achieved with regards to the Adaptive PoPA expressed as follows:

$$J_{Pinch} = \min_{U_c} \sum_{k=1}^{N-1} f(\varepsilon_i(k)\,,\, SOAcc_l^m(k), U_c(k)) \tag{1}$$

Subject to the Power Pinch analysis constraints:

$$S_{Lo}^l \le SOAcc_l^m(k) \le S_{Up}^l \tag{2}$$

$$SOAcc_l^n(k_1) \cong SOAcc_l^m(N) \tag{3}$$

$$\varepsilon_{EL}(k) + \varepsilon_{FC}(k) \le 1 \tag{4}$$

where, k_1 is the first hour , ε_i (t) is a binary variable for the dispatchable asset's state $i \in [FC, EL]$, (see appendix I), U_c (k) represents the PoPA EMS control variable and subscript $c \in \{FC, EL\}$ indicates the dispatchable asset. In $SOAcc_l^{m,n}$ the superscripts m and n refers to the predicted and real $SOAcc$ respectively, and subscript $l \in \{BAT, HT, WT\}$ indicates the energy storage of note.

The constraints imposed by (2) ensures the pinch operating limits are not violated. The duty cycle of the energy storage is preserved by the terminal constraint (3) to infer the available energy at the end of the prediction horizon N (AEEND). The binary variable constraint (4) prevents the simultaneous dispatch of assets that concurrently consume and produce the same energy carrier (e.g. FC and EL).

The following explanation is for one asset, the BAT, but is relevant to all asset types. At every time step k, the proposed algorithm compares the forecast and real $SOAcc^{n}_{BAT}(k)$ for inconsistency or forecast deviation via a state feedback close loop [72]. As illustrated in Figure 4a, ΔH exceeds 5% at time $k+2$. Therefore, state correction is effected at the next time $k+kmin$, to decrease the forecast deviation between the predicted $SOAcc^{m}_{BAT}$ and actual $SOAcc^{n}_{BAT}$. The re-computation of the PGCC (dotted black line in Figure 4a) which follows reveals an anticipated violation of the S_{UP} such that $SOAcc^{m}_{BAT}$ is a maximum at time $k+11$, and the AEEND. Thus, the predicted PGCC is re-shaped as shown in Figure 4b (blue line) with the EL dispatched at time $k+10$ and $N-1$.

The error $e(k)$ and magnitude of uncertainty ΔH between the forecast and real state of charge of the Battery are expressed in (5) and (6) respectively as follows:

$$e(k) = SOAcc^{n}_{BAT}(k) - SOAcc^{m}_{BAT}(k|k-1) \tag{5}$$

$$\Delta H(k) = |e(k)| \tag{6}$$

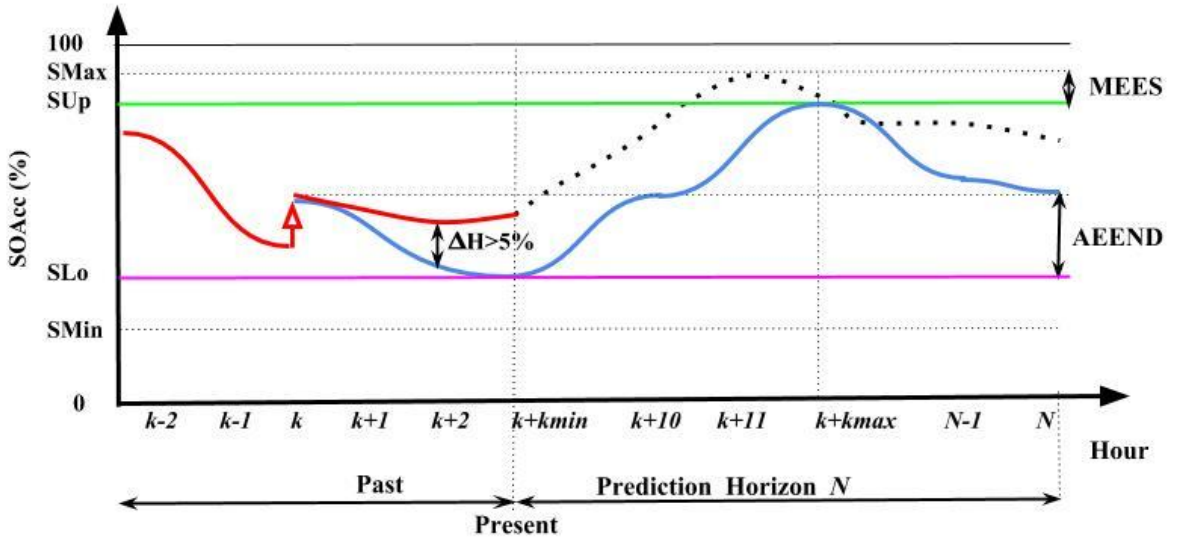
where, $SOAcc^{m}_{BAT}(k|k-1)$ is the predicted battery state of charge at time k based on a prior time step $k-1$ and $SOAcc^{n}_{BAT}(k)$ is the actual battery state of charge at time step k.

Furthermore, if ΔH is greater than the deviation threshold ξ at any sampling instance, the PoPA is repeated in the predictive horizon in order to determine the optimal dispatch and schedule sequence from that instant up until time N. ξ (which may be varied or decreased for a tighter bound) is set at 5%, to ensure minimal forecast deviations as well as to reduce any computational cost. Re-computation of the PGCC uses equations (7) - (8) as follows:

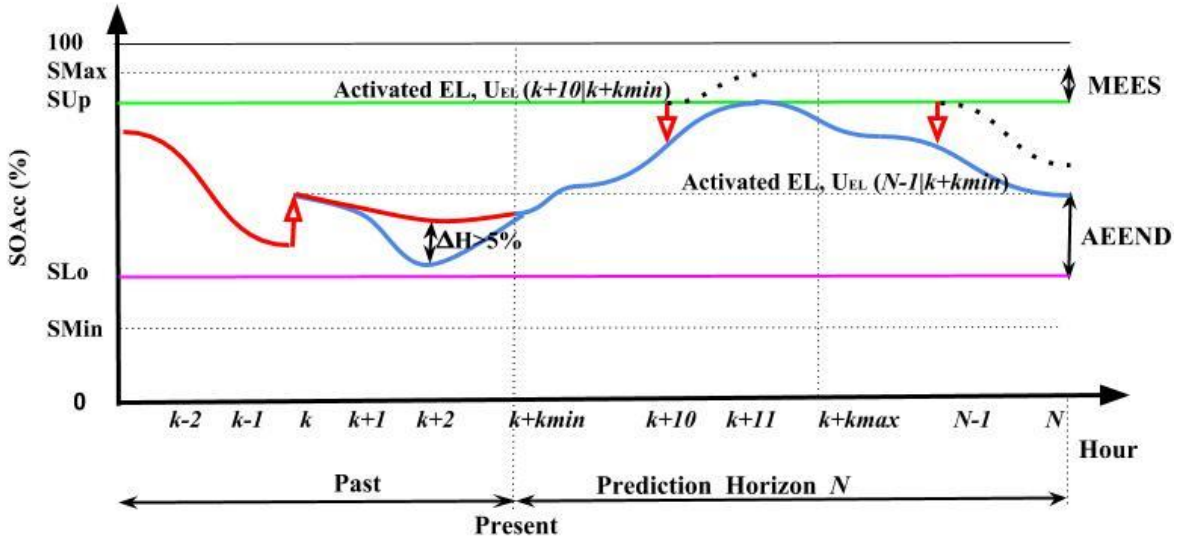
$$SOAcc^{m}_{BAT}(k) := \begin{cases} f\big(\Delta H(k)\big) & if\ \ \Delta H(k) > \xi \\ SOAcc^{m}_{BAT}(k|k-1) & Otherwise \end{cases} ,\forall_k \tag{7}$$

Where, $f\big(\Delta\mathcal{H}(k)\big)$ corrects $SOAcc^{m}_{BAT}$ as follows:

$$f\big(\Delta H(k)\big) = \begin{cases} SOAcc^{m}_{BAT}(k|k-1) + \Delta H(k) & e(k) > 0 \\ SOAcc^{m}_{BAT}(k|k-1) - \Delta H(k) & e(k) < 0 \end{cases} \tag{8}$$



(a)



(b)

Fig. 4. (a) State error correction and (b) re-shaped PGCC with Adaptive PoPA

4. Kalman Filter Adaptive Power Pinch Analysis

In the previous section a reactive error correction strategy has been presented, the adaptive PoPA, which does not consider the effect of future un-modelled uncertainty. This may result in a limit violation as shown in Figure 5a. Therefore, the Kalman filter is incorporated into the Adaptive PoPA framework for robustness, as the battery's future state ($SOAcc^{m}_{BAT}(k+1|k)$ is predicted while incorporating the effect of uncertainty at each time interval

upon the availability of the most recent battery state ($SOAcc^{n}_{BAT}(k)$) measurement. In order to predict the battery's state, a priori error covariance $\mathcal{P}_{k-1}$ matrix with respect to $SOAcc_l$, updates the Kalman gain $K_{G(k)}$ as follows:

$$K_{G(k)} = \mathcal{P}_{k-1}\,\mathcal{J}^T\,[\mathcal{J}\,\mathcal{P}_{k-1}\,\mathcal{J}^T + \mathcal{R}_k]^{-1} \tag{9}$$

The updated Kalman gain is used to update the a priori covariance matrix:

$$\mathcal{P}_k = \left[\mathcal{J} - K_{G(k)}\mathcal{J}\right]\mathcal{P}_{k-1} \tag{10}$$

The most recent output state measurement $SOAcc^{n}_{l}(k)$ is used to update the estimated state as follows:

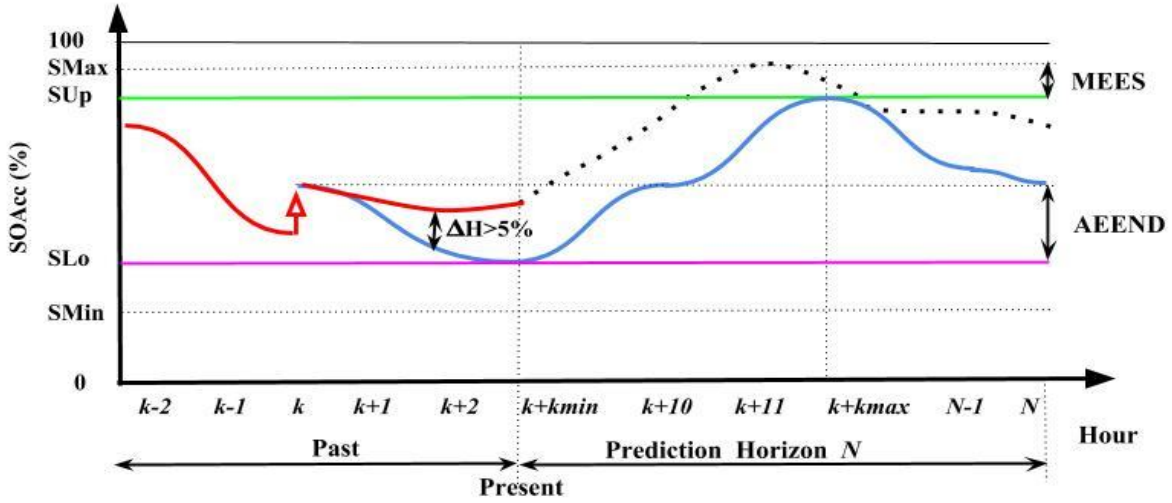
$$SOAcc^{m}_{l}(k) = SOAcc^{m}_{l}(k|k-1) + K_G(SOAcc^{n}_{l}(k) - \mathcal{J}_k SOAcc^{m}_{l}(k|k-1)) \tag{11}$$

The posterior error covariance matrix is also updated:

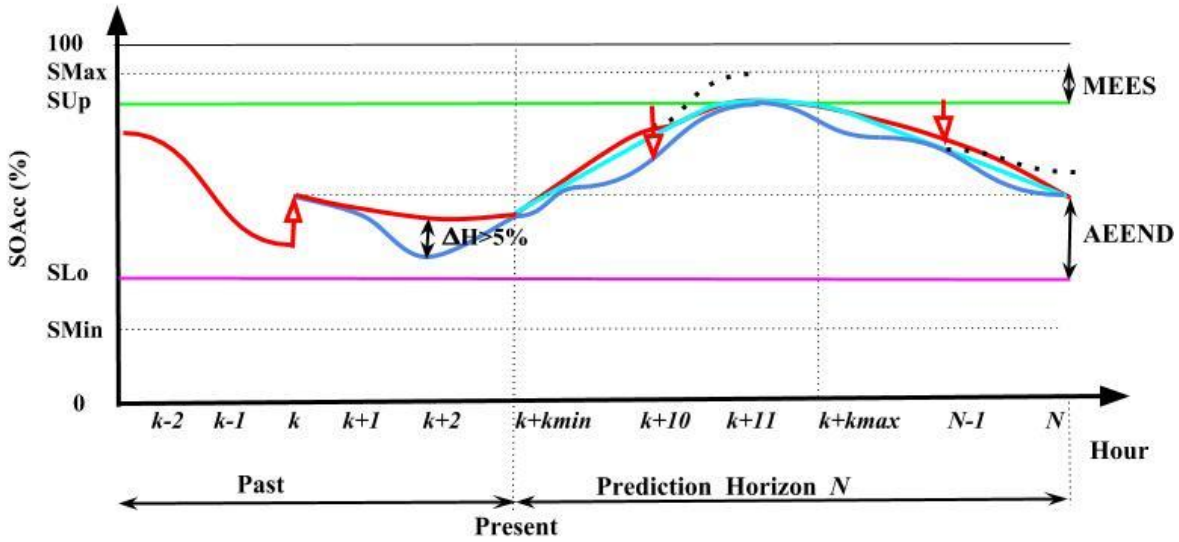
$$\mathcal{P}_{k+1} = \mathrm{A}\,\mathcal{P}_k\,\mathrm{A}^T + \mathcal{R}_k \tag{12}$$

Where, $\mathrm{A} \in l$ x l is an identity state transition matrix for the energy storages l, $\mathcal{J}_k \in$ l x l is an identity matrix and $\mathcal{R}_k$ is the covariance noise matrix related to the uncertainty in $SOAcc^{m}_{l}$.

Therefore, this formulation can be used to consider a multi-vector case of uncertainty in the energy storages. Nevertheless, in this work only the $SoAcc$ of the BAT is the parameter directly impacted by the LD and RES uncertainty since it acts as the central integrating ES, and a change in the $SoAcc$ of HT and WT can be considered deterministic as well as contingent on the controlled activation of FC or EL. Therefore, the variance and co-variance of $SoAcc$ of HT and WT in $\mathcal{P}_k$ matrix are set to 0. Furthermore, the $SOAcc^{m}_{BAT}(k) \in [SOAcc^{m}_{l}(k)]$ is determined in (11) in order to identify the uncertainty over successive k- steps ahead and consequently to compute the PGCC. Thereafter, the PGCC is re-shaped via PoPA minimum energy targeting as before. Thus, a sequence of dynamic EMSs which satisfies both the PoPA S_{LO} and S_{UP} constraints with uncertainty projection is realised in the prediction horizon for the optimal dispatch and scheduling of energy resources in the control horizon. The concept is illustrated in Figure 5b, where the cyan plot indicates the PGCC re-shaped via the Kalman+Adaptive PoPA. The violation of the S_{UP} at time $k + 11$, which occurred with the Adaptive PoPA EMS in Figure 5a, is avoided by dispatching the EL to recover correct MESS at time k+10. Likewise, the procedure is repeated for the AEEND constraint. Figure 6, shows the Kalman+Adaptive PoPA algorithm.



(a)



b)

Fig. 5. (a) PGCC shaped with Adaptive PoPA and (b) PGCC shaped with Kalman+Adaptive PoPA

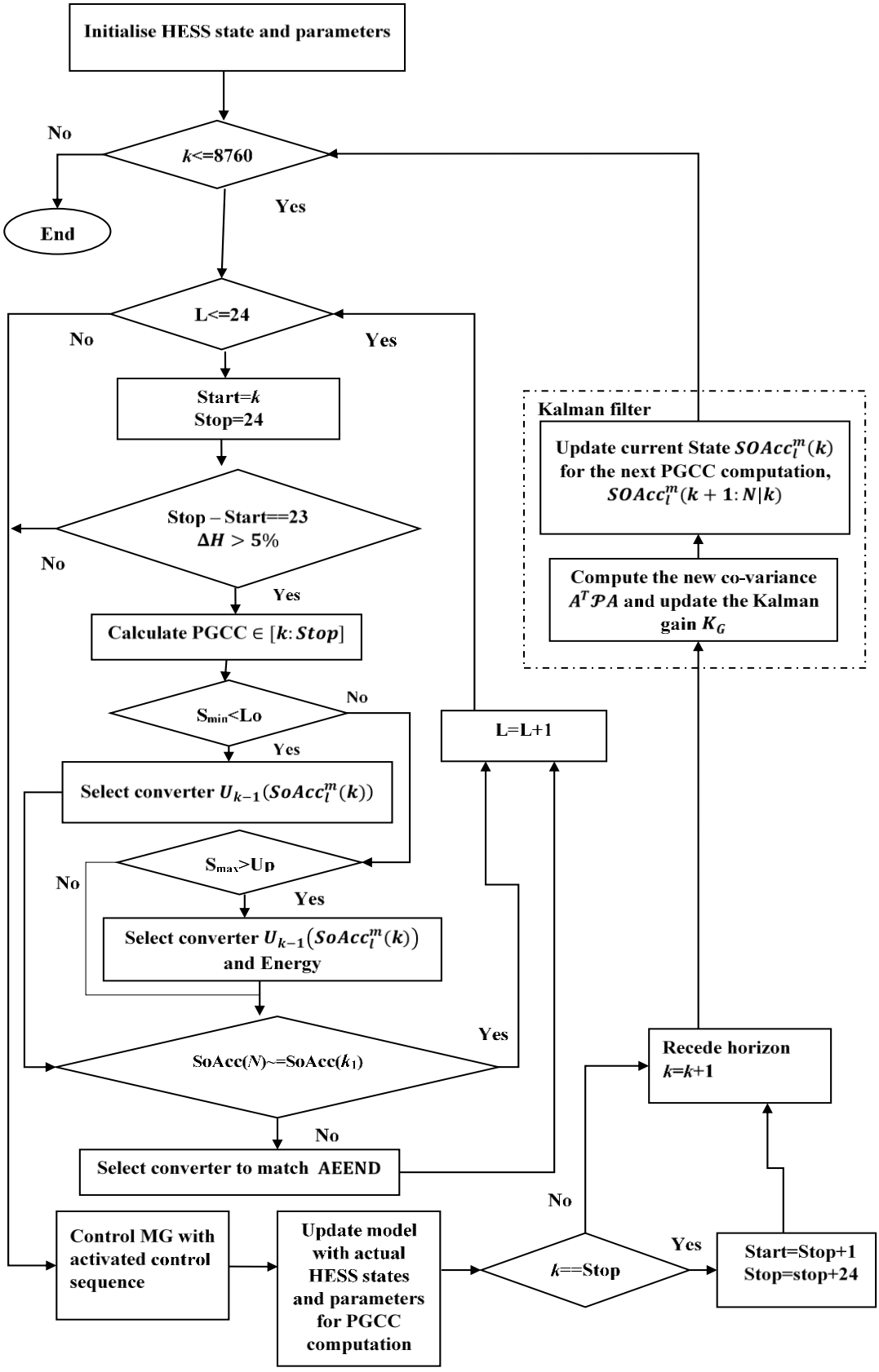


Fig. 6. Kalman +Adaptive Power Pinch Algorithm

5. Reinforcement Learning Adaptive Power Pinch Analysis

The approach presented in this work involves formulating the uncertainty problem as a MDP considered in the discrete time step k, where an agent has to act optimally by inferring an action in each state as determined by the adaptive MPC PoPA trajectory.

The MDP is a tuple(S, A, R, S', A') where:

$\mathcal{S}$: is a set of discrete *n*-states $\mathcal{S} = \{s_1, s_2, \ldots, s_n\}$ and s_k denotes the state of the environment at time step *k*.

In this work, $s_k := f\langle SOAcc_{BAT}^{m}(k), SOAcc_{BAT}^{n}(k),\ e(k)\rangle$ (13)

$\mathcal{A}$: is a discrete set of *n*-actions for selection by the agent $\mathcal{A} = \{a_1, a_2, \ldots, a_7\}$ and a_k indicates the selected action at time *k*.

Furthermore, the set of dispatchable assets for the PGCC shaping is expressed as follows:

$U_c(t) \subseteq \mathcal{A}_k := \{a_1,\ \delta_1 FC, \delta_2 FC, \delta_3 FC, \delta_4 EL, \delta_5 EL, \delta_6 EL\}$

Where, $\delta_x,\ x \in [1{:}\,6]$, represents percentage proportions {10, 50, 90} and {10, 50, 100} of corresponding flow of energy/material $F_{FC\to BAT}^{Pow}(k)$ and $F_{BAT\to EL}^{Pow}(k)$ respectively to a selected action and a_1 denotes null action.

$\mathcal{T}(s, a, s')$: is the probability of transitioning to a next state s' from state s over a given set of transitions when an action a is chosen.

$\mathcal{S}\ x\ \mathcal{A} \to\ R$: An immediate reward r_t is received as a result of the system state transition $\mathcal{T}(s, a)$ to the next state s' by mapping state and action pair (s, a) due to a decision making policy π.

Therefore, both the transition and reward probability distributions are implicitly Markov properties where the future state s' only depends on the present state s. The current action a is independent of the past state(s) s^- that lead to the present state [74, 75].

$\mathcal{T}(s'|s^-, s, a) = \mathcal{T}(s'|s, a)$ (14)

The model of the system is required for initial training of the agent in order to infer the control action on the actual system from the MPC-PoPA. The agent adapts to the real system over time and retrains on newer samples. The MDP learning agent learns the optimal policy $\pi^*(a|s)$ from accumulated past experience which maps an optimal action to a given state. Hence, this maximises the cumulative scalar reward return as shown in (15).

$$\mathcal{V}^{\pi} = E\left[\sum_{k=1}^{\infty} \gamma^{k-1} r_k(s_1, a_1|\pi)\right] \quad (15)$$

The Q-function Q^{π}(s, a) for a given MDP represents the optimal value function $\mathcal{V}^{\pi*}$.

The agent learns the optimal action to take in the environment through experience by taking actions in the environment while learning the optimal policy.

The Q-learning rule after taking an action a in a state *s*, obtaining a reward r and transitioning to s' is as follows:

$$Q_k(\mathrm{s},\mathrm{a}) = \begin{cases} Q_k(\mathrm{s},\mathrm{a}) + \alpha\ [r_k + \gamma \max_{a'} Q_{k+1}(\mathrm{s}',\mathrm{a}') - Q_k(\mathrm{s},\mathrm{a})] & \forall\ k = [1,2,\ldots N-2] \\ Q_k(\mathrm{s},\mathrm{a}) + \alpha\ [r_k - Q_k(\mathrm{s},\mathrm{a})] & \forall\ k = N-1 \\ Q_k(\mathrm{s},\mathrm{a}) & \forall\ k = N \end{cases} \quad \alpha, \gamma\ \in [0, < 1] \quad (16)$$

Where α, γ are learning rate and future reward discount factor with the future discounted reward omitted during the update of the agent at a terminal state at time step $N-1$.

5.1 Planning stage for the Q-learning Agent

The MPC-PoPA model is used to bootstrap the Q-learning agent to ensure that the agent acts optimally with respect to tracking the PoPA trajectory, computed offline prior to online deployment so as to minimise and avoid exploiting costly mistakes on the real system. The advantage of the Q-algorithm is that the agent garners experience from the real environment and retrains offline by replaying the experience after each episode at time N to further reinforce the learning agent's Q - value to guarantee optimality. The model-free learning happens using the Q-learning algorithm and switches to a Monte Carlo algorithm at $N-1$ which denotes the terminal state (horizon) for the agent, as shown in (16). Therefore, the learning involves two steps; a direct and indirect learning, from the model and from the actual system (environment) respectively.

5.2 Action Selection

The action selection approach in (17) which has been modified to include safety precautions in critical states (near the Pinch limits), is based on the probability (1 - θ) of selecting a $greedy$ policy $\pi(\mathrm{s})$ over a random action with probability of θ [76, 77]. This approach exploits the best action as indicated by the maximum value function $Q^{\pi*}(s,a)$ for a given state while performing exploration with the inverse probability (θ) of acting greedily. This strategy strikes a balance between exploration and exploitation while satisfying the famous Bellman's principle of optimality [78], minimizing the deviation of the system controlled by the learning agent from the Pinch target, and exploring the state space. If the $SOAcc_{BAT}^{n}(k)$ is less than Lo or greater than Up, the FC and EL are dispatched by the agent respectively. Furthermore, the AEEND constraint imposed at the end of the day is achieved by overriding the agent's action with the Adaptive PoPA's EMS. The action policy $\pi(\mathrm{s})$ is expressed as follows:

$$\pi(\mathrm{s})=\begin{cases} a_k(s) & If\ U < \ greedy\ action\ probability\ (1-\theta) \\ \delta_3 FC & if\ U > \ greedy\ action\ probability\ (1-\theta) \wedge SOAcc_{BAT}^{n}(k) \leq 30\% \\ \delta_6 FC & if\ U > \ greedy\ action\ probability\ (1-\theta) \wedge SOAcc_{BAT}^{n}(k) \geq 90\% \\ select\ a\ random\ action & Otherwise \end{cases} \tag{17}$$

Where,

U is a randomly generated value between 0 and 1 given each k time step.

$$a_k(s) := \begin{cases} \delta_3 FC & SOAcc_{BAT}^{n}(k) \leq 30\% \\ \delta_6 EL & SOAcc_{BAT}^{n}(k) \geq 90\% \\ \underset{a_k(s) \subseteq \{a_1, \delta_n FC\}, n \in [1:3]}{argmax} Q(s_k, a_k) & SOAcc_{BAT}^{n}(k) \geq 30\% \wedge SOAcc_{BAT}^{n}(k) \leq 40\% \\ \underset{a_k(s) \subseteq \{a_1, \delta_n EL\}, n \in [4:6]}{argmax} Q(s_k, a_k) & SOAcc_{BAT}^{n}(k) \geq 80\% \wedge SOAcc_{BAT}^{n}(k) \leq 90\% \\ \underset{a_k(s) \subseteq A_t}{argmax} Q(s_k, a_k) & otherwise \end{cases} \tag{18}$$

5.3 Reward Function Formalisation

In order to train the Q-learning agent, a suitable reward function is expressed mathematically. This is such that the agent follows the optimal policy $\pi^*(s)$ which minimises the cost function between the agent's off-policy and the adaptive MPC PoPA trajectory, and is expressed as follows:

$$J_\pi\left(SOAcc_{BAT}^{n}\right) = \lim_{k \to N-2} E\left[\sum_{k=1}^{N-2} |SOAcc_{BAT}^{m} - SOAcc_{BAT}^{n}|^2 + \left(\gamma J_\pi\left(s_{k+1}\right)\right)\right] \tag{19}$$

Thus, it follows that:

$$\min_{U_c} J_\pi\left(SOAcc_{BAT}^{n}\right) \triangleq \lim_{k \to \infty} \underset{a_k \in A_k}{argmax} E\left[\sum_{k=N-2}^{\infty} \left(\gamma^{k-1} \mathcal{R}(s_{k+1}, a_{k+1})\right)^{-1}\right] \tag{20}$$

The reward function in (21) is aimed at accelerating learning. It comprises of a fixed reward G, with penalty factors W_1 and W_2, representing a squared error penalty cost function and constant penalty factor respectively.

The magnitude of the W_1 penalty factor is such that it increases proportionally to the absolute squared error deviation from the pinch target at that instant and the systems state if the agent takes a suboptimal action as shown in equation (22). Furthermore, the rewarded function in (23) - (25) is able to update the agent $Q(s, a)$ regardless of whether the availability proposition $\varepsilon_i^{Avl}(k)$ (see appendix II) for both the FC and EL assets are met, while exploiting an action which minimises the error cost.

A typical illustration; if the operating point dictated by Adaptive PoPA anticipates future energy deficit and requests activation of the FC, while the agent activates the EL, a penalty would suffice. Thus, the penalty function, serves as a closed loop negative feedback to the agent. Therefore, in order to obtain the maximum reward G at a given time step, the action performed by the agent, must satisfy the consequent conditional proposition. Thus, if a_k results in $SOAcc_{BAT}^{n}(k+1)$ being greater than or equal to $SOAcc_{BAT}^{m}(k+1)$ and the agent's action a_k, is equal to the optimal action Uc_{min} the maximum reward G, is obtained. As shown in (23) Uc_{min} is contingent on function D and E in equation (24) and (25) respectively. Where, functions D and E are performed abstractly by iterating over all actions a_i the agent can perform if the $SOAcc_{BAT}^{m}(k+1)$ is greater than 80% and less than 80% respectively. Specifically, assuming the $SOAcc_{BAT}^{m}(k+1)$ is less than 80%, function D is used and thus by

iterating over all actions $a_i\ i \in [1:7]$, Uc_{min} becomes the minimum (infimum) action which results in $SOAcc^m_{BAT}(k+1)$ being greater or equal to $SOAcc^n_{BAT}(k+1)$. This supresses the excessive usage of the FC. Similarly, where function E suffices, the maximum (supremum) action which results in $SOAcc^m_{BAT}(k+1)$ being less than or equal to $SOAcc^n_{BAT}(k+1)$ becomes Uc_{min} such that the EL is used optimally.

Furthermore, if the action performed by the agent is not equal ($\neg=$) to Uc_{min}, and consequently $SOAcc^n_{BAT}(k+1)$ becomes less than or equal to $SOAcc^m_{BAT}(k+1)$ a negative penalty denoted by $-W_1$ ensues in other to apprise the agent from exploiting adverse actions which over discharges the BAT.

Also, where the agent performs a_k not equal to Uc_{min}, but which results in the $SOAcc^n_{BAT}(k+1)$ becoming greater than or equal to $SOAcc^m_{BAT}(k+1)$, a penalty W_1 is deducted from the maximum reward G in order to dampen excessive usage of the FC. Similarly, a penalty $-(W_1+W_2)$ is used to accelerate the agent's learning curve if successive violations of any of the pinch limits occur as a result of suboptimal action.

The reward function proposition for $\mathcal{S}\ x\ \mathcal{A} : \mathcal{R}(\mathcal{S},\mathcal{A})$ is implemented as follows;

$$\mathcal{R}(s_k,a_k)=\left\{\begin{array}{ll} G & \left[\begin{array}{c}[SOAcc^n_{BAT}(k+1)\geq SOAcc^m_{BAT}(k+1)\wedge a_k == Uc_{min}\ \wedge \\ [SOAcc^n_{BAT}(k+1)>S^l_{Lo}\wedge\ SOAcc^n_{BAT}(k+1)<S^l_{Up}]\end{array}\right] \\ -W_1 & \left[\begin{array}{c}[SOAcc^n_{BAT}(k+1)\leq SOAcc^m_{BAT}(k+1)]\wedge a_k \neg= Uc_{min}\ \wedge \\ [SOAcc^n_{BAT}(k+1)>S^l_{Lo}\wedge\ SOAcc^n_{BAT}(k+1)<S^l_{Up}]\end{array}\right] \\ G-W_1 & \left[\begin{array}{c}[SOAcc^n_{BAT}(k+1)\geq SOAcc^m_{BAT}(k+1)]\wedge\ \ a_k \neg= Uc_{min}\ \wedge \\ [SOAcc^n_{BAT}(k+1)>S^l_{Lo}\wedge\ SOAcc^n_{BAT}(k+1)<S^l_{Up}]\end{array}\right] \\ -(W_1+W_2) & \left[\begin{array}{c}[SOAcc^n_{BAT}(k)\leq SOAcc^n_{BAT}(k+1)]\wedge \\ [SOAcc^n_{BAT}(k)\geq S^l_{Up}\wedge SOAcc^n_{BAT}(k+1)\geq S^l_{Up}]\wedge \\ a_k \neg= Uc_{min}\vee\ SOAcc^n_{BAT}(k+1)\geq S^l_{Up}\wedge a_k \neg= Uc_{min}\end{array}\right] \vee \\ & \left[\begin{array}{c}[SOAcc^n_{BAT}(k)\leq SOAcc^n_{BAT}(k+1)]\wedge \\ [SOAcc^n_{BAT}(k)\geq S^l_{Up}\wedge SOAcc^n_{BAT}(k+1)\geq S^l_{Up}]\wedge \\ a_k \neg= Uc_{min}\ \vee\ SOAcc^n_{BAT}(k+1)\leq S^l_{Lo}\wedge\ a_k \neg= Uc_{min}\end{array}\right] \end{array}\right\} \tag{21}$$

Where, $W_1\ and\ W_2$ are penalty factors for reward shaping.

$$W_1=[(SOAcc^n_l(k+1)-SOAcc^m_{BAT}(k+1))/SOAcc^m_{BAT}(k+1)]^2 \tag{22}$$

The action which results in the minimum optimal control action is derived abstractly as follows:

$$Uc_{min}\coloneqq\begin{Bmatrix} D & SOAcc^m_{BAT}(k+1)>S^l_{Lo}\wedge\ SOAcc^m_{BAT}(k+1)\leq(S^l_{Up}-10\%) \\ E & SOAcc^m_{BAT}(k+1)>(S^l_{Lo}+50\%)\wedge\ SOAcc^m_{BAT}(k+1)<(S^l_{Up}) \end{Bmatrix} \tag{23}$$

Where,

$$D := \inf\{(SOAcc_{BAT}^{m}(k+1) | \sum_{i=1}^{7} Q(a_i, s_{k+1})) \geq SOAcc_{BAT}^{n}(k+1)\} \quad (24)$$

$$E := \sup\{(SOAcc_{BAT}^{m}(k+1) | \sum_{i=1}^{7} Q(a_i, s_{k+1})) \leq SOAcc_{BAT}^{n}(k+1)\} \quad (25)$$

During the online deployment, the PoPA target is modified respectively with the MOES or MEES so as to capture the effect of uncertainty after S_{LO} and S_{UP} violation occurs at any instant as follows:

$$SOAcc_{BAT}^{m}(k|k) := \begin{cases} S_{Up}^{l} & SOAcc_{BAT}^{n}(k) > S_{Up}^{l} \\ S_{Lo}^{l} & SOAcc_{BAT}^{n}(k) < S_{Lo}^{l} \end{cases}, \quad \forall_t \ if \ \exists \ \Delta H(k) \neq 0 \quad (26)$$

The reward function is modified to incorporate the MOES or MEES thus guaranteeing the model-free agent will act optimally in the event of uncertainty to maximise the expected reward:

$$J_{Pinch}(SOAcc_{BAT}^{n}) + J_e(\Delta H) = \min_{U_c} J_\pi(SOAcc_{BAT}^{n}) \quad (27)$$

Furthermore, by performing the optimal policy π^* the corresponding cost is as follows:

$$J_\pi^*(SOAcc_{BAT}^{n}) \rightarrow \lim_{k \rightarrow \infty} E\left[\sum_{k}^{\infty} \gamma \left(J_{Pinch}(SOAcc_{BAT}^{n}) + J_e(\Delta H)\right)\right] \quad (28)$$

Since the cost of the error due to uncertainty tends to zero when following the optimal policy, J_π^*(s), the agent incorporates the uncertainty estimation into the PoPA:

$$\lim_{k \rightarrow \infty} \quad J_{\pi(k)}^*(SOAcc_{BAT}^{n}) \leq \gamma J_{Pinch(k)}(SOAcc_{BAT}^{n}) \quad (29)$$

The expected cost following the pinch analysis and uncertainty propagation is less than following only the PoPA model. Hence, the experience of the agent integrated into the MPC Adaptive PoPA framework guarantees optimal operation, as long as the conditions of optimal action selection and learning rate decay are satisfied. Figure 7 and 8, illustrates the RL+Adaptive PoPA architecture and algorithm respectively. Furthermore, the pseudo codes for the proposed algorithms are presented in Appendix I.

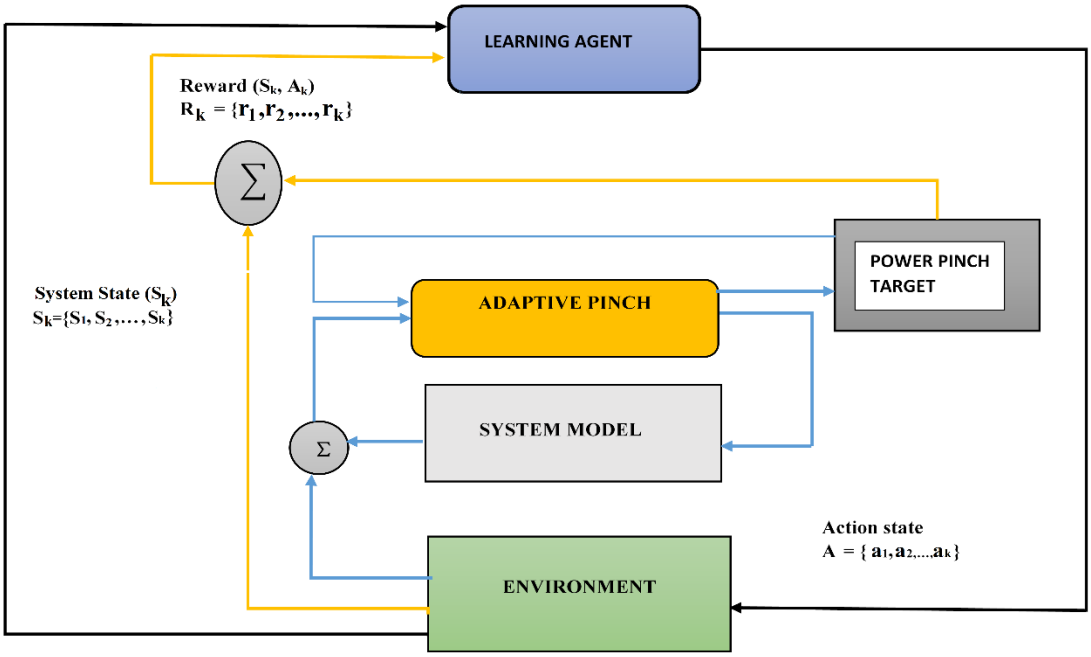


Fig. 7. Reinforcement Learning Adaptive Power Pinch architecture

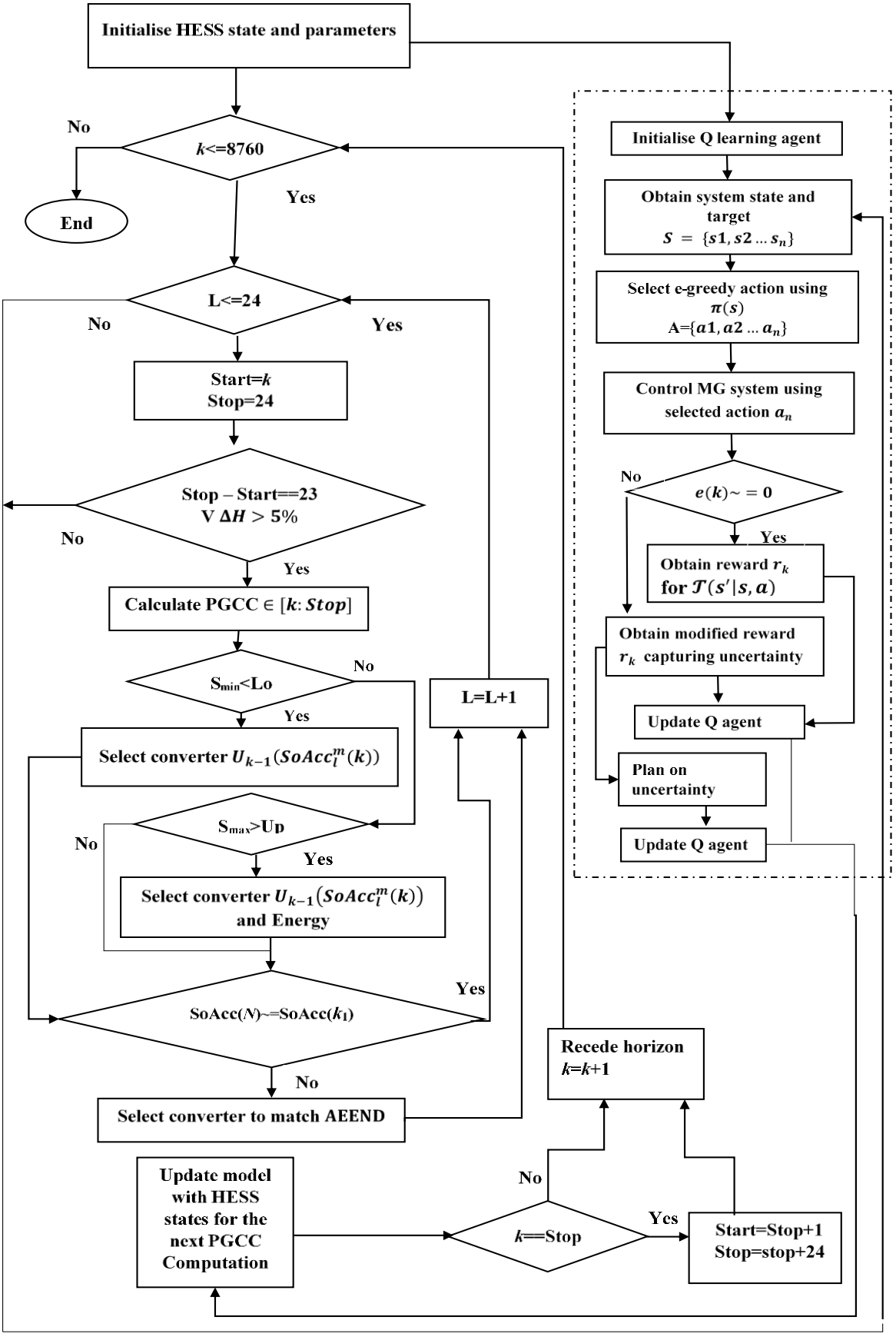


Fig. 8. RL + Adaptive Power Pinch Algorithm

6. Results and Discussion

The three new methods are evaluated against the DA-PoPA in a short (three days (72h)) and long-term (one year (8760 h)) deployment in a stand-alone HESS. The initial conditions for the $SOAcc_l^m$ is such that $l \in \{\text{BAT, HT and WT}\}$ corresponds to 70%, 80% and 30% respectively. The HESS parameters used as case study are derived from an existing real system [9] as shown in Table 1. Also, real load demand profiles for a typical residential home and solar irradiance data pertaining to Newcastle, United Kingdom, are sourced from ELEXON [79] and NREL [80] respectively.

Table 1

HESS Micro-grid parameters [9]

System Components	Specification
Load (peak)	2200 W
PV (66.64 W rated power)	217
DSL	2210 W
BAT	3000 Ah / 48 V
FC	3000 W
EL	4000 W
HT	30 bar, 15 m^3
$\eta_{CV}, \eta_{PV}, \eta_{FC}, \eta_{EL}$	0.95, 0.10, 0.87, 0.87

The performance main indices (30) - (32) used in evaluating the EM approaches are with respect to the total number of times the S_{Lo}^l (30%) and S_{Up}^l (90%) Pinch limits are violated and the DSL activated, as follows [42];

$$\text{Sum of Deficit}=\sum_{k=1}^{N=8760}\begin{cases}1 & S_{Lo}^l > SOAcc_{BAT}^n(k) \\ 0 & otherwise\end{cases} \quad (30)$$

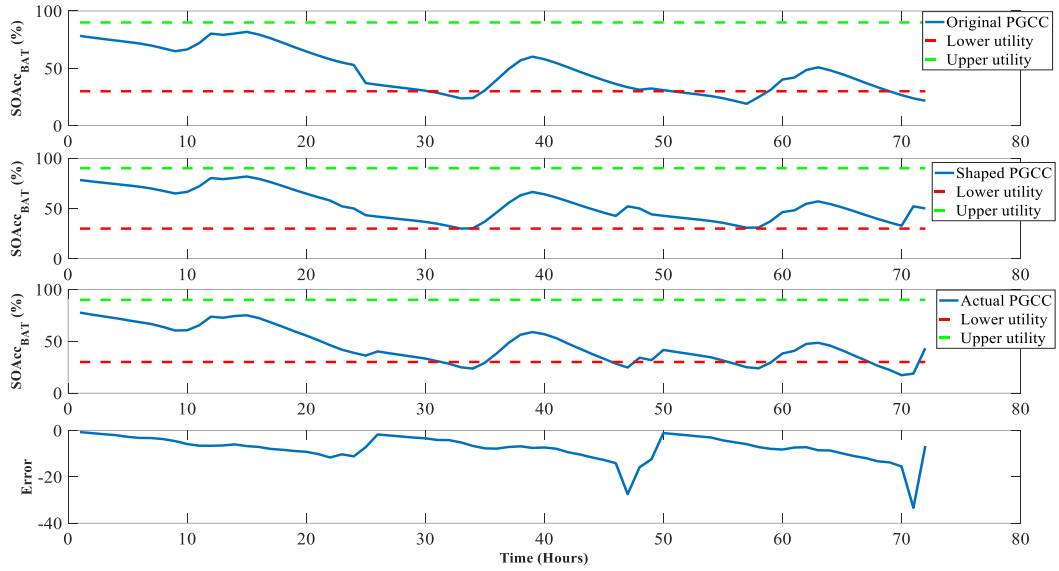
$$\text{Sum of Surplus}=\sum_{k=1}^{N=8760}\begin{cases}1 & S_{Up}^l > SOAcc_{BAT}^n(k) \\ 0 & otherwise\end{cases} \quad (31)$$

$$\text{Sum of DSL activation}=\sum_{k=1}^{N=8760}\begin{cases}1 & 20\% > SOAcc_{BAT}^n(k) \\ 0 & otherwise\end{cases} \quad (32)$$

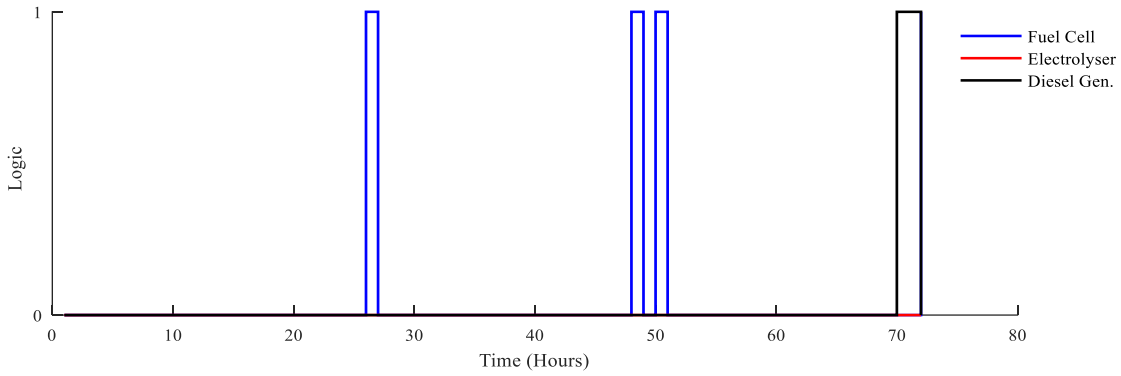
6.1 Short-term operation

6.1.1 Day – Ahead Power Pinch Analysis

As illustrated in Figures 9(a), the original PGCC show the $SOAcc_{BAT}^m$ would dip successively below the S_{LO} due to impending energy deficit within the first 72 h, if electricity is not outsourced in advance. Thus the PGCC is shaped accordingly by activating the FC four times as shown in Figure 9 (b). However, the PGCC continuously violated S_{LO} 14 time instances which led to the activation of the DSL twice due to uncertainty indicated by the error plot as shown in Figure 8a, regardless of hydrogen availability.



a)

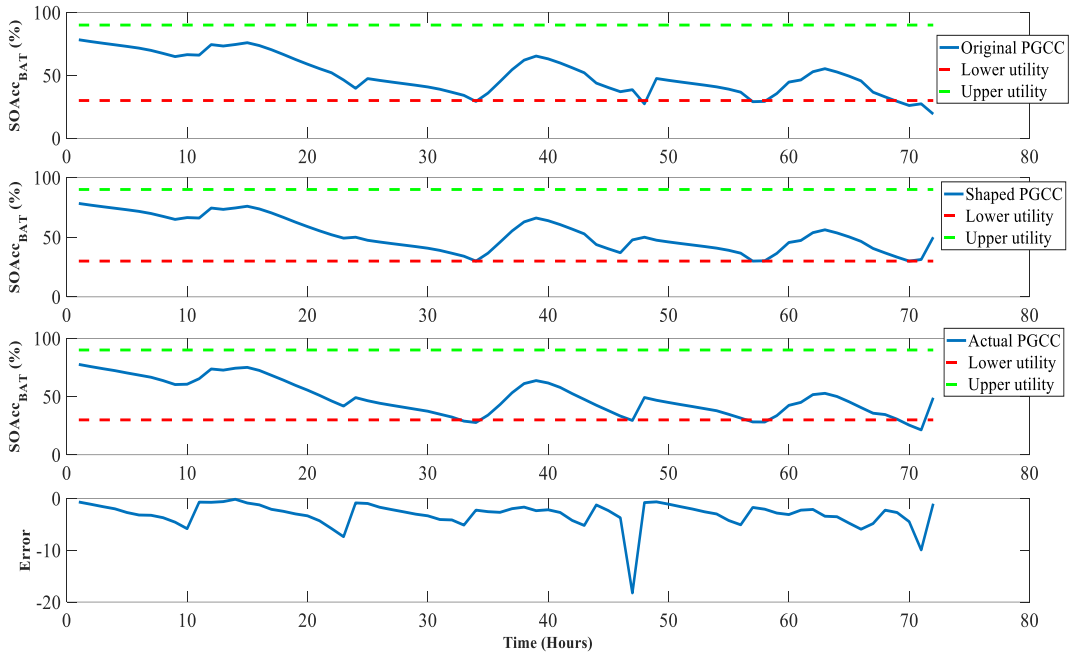


b)

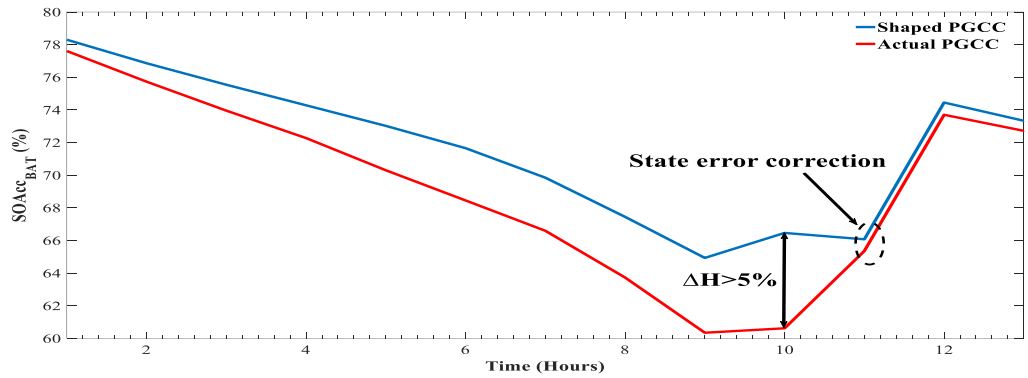
Figure 9: a) DA-PoPA response and b) Dispatchable Logic state for the first 72h of the year

6.1.2 Adaptive Power Pinch Analysis Energy Management Strategy for Uncertainty

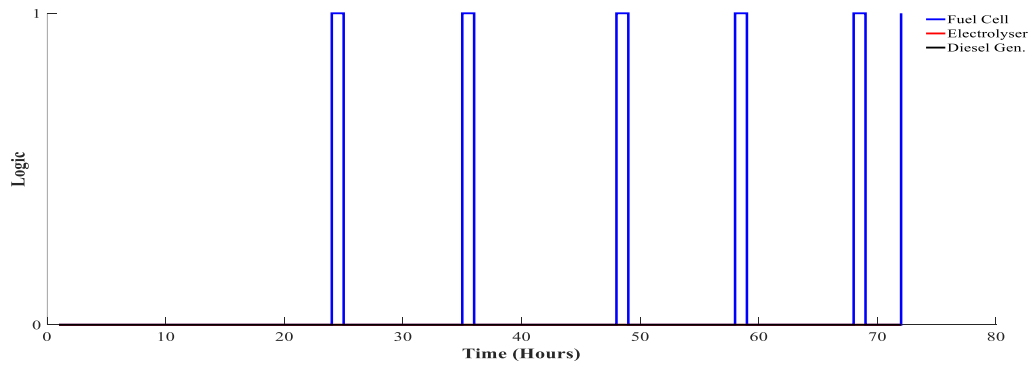
The energy deficit and consequent forecast error deviation exhibited by the DA-PoPA was reduced by the dynamic shaping of the PGCC within a receding control horizon as shown in Figure 10(a). Figure 10(b) illustrates the state error correction at the inception of the 11:00 Hr after ΔH became greater than 5% at 10:00 h. However, the $SOAcc_{BAT}^{n}$ dipped at the 33rd, 34th, 47th, 57th, 58th, 70th, and 71st h, without activating the DSL. Furthermore, despite dispatching the FC six times, as shown in Figure 10(c) after the occurrence of the unforeseen dip, a further violation of S_{LO} re-occurred. This was because the MOES delivered by the FC was less than required, due to deficit energy target variability. The successive dips underscore the need for a preventive approach since the reactive approach only responds after the forecast error has occurred.



(a)



(b)

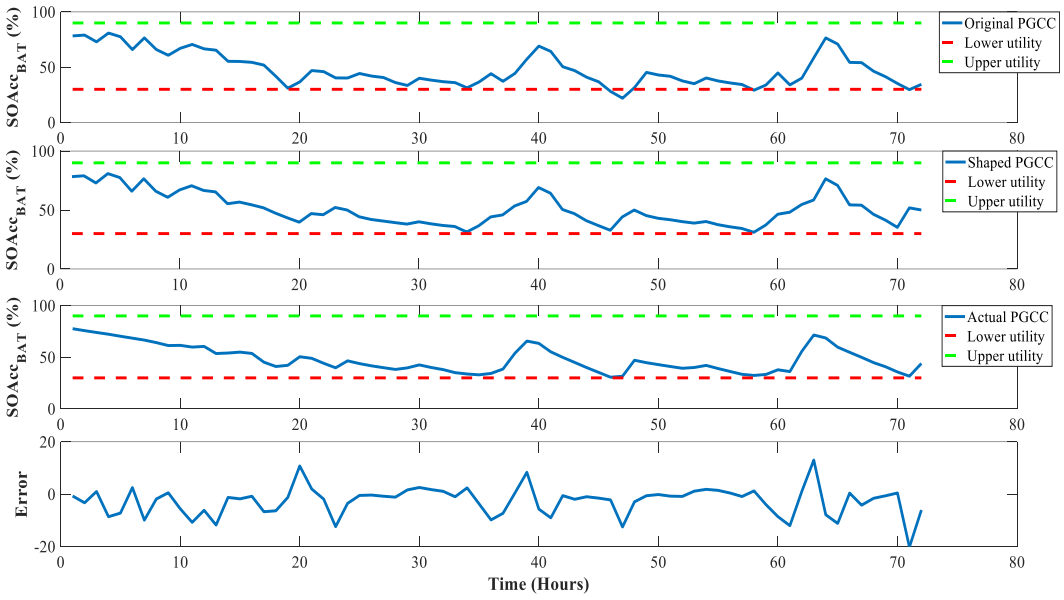


(c)

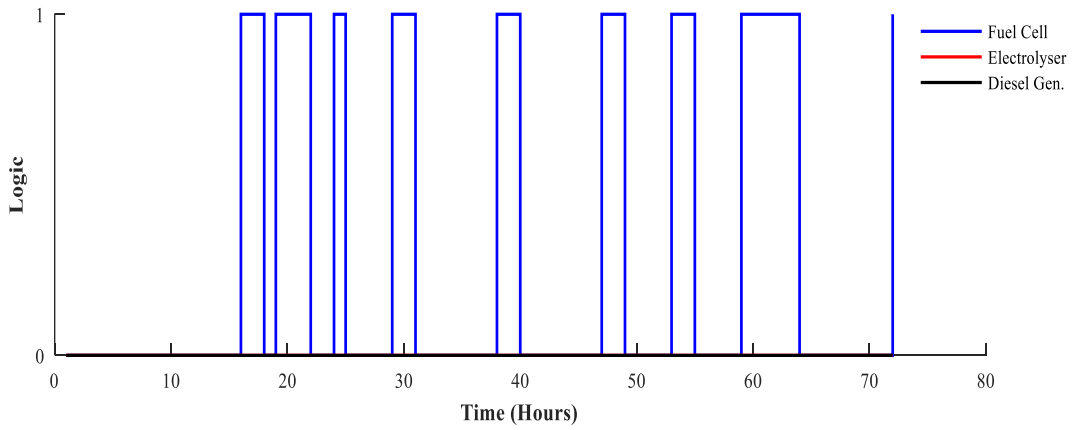
Fig. 10. a) Adaptive Power PoPA, b) State error correction and c) Converter Logic

6.1.3 Kalman Filter Adaptive PoPA

The Kalman + Adaptive approach results in the PGCC violating S_{LO} 7 times at time 49:00 - 56:00 h and at time 64:00 - 70:00 h, as shown in Figure 11a. Additionally, the FC was activated 20 times in response to uncertainty with the DSL never activated as shown in Figure 11 (b). The Kalman+Adaptive PGCC closely matched the actual state of the plant as shown in Figure 11(a), with the uncertainty adequately propagated within the first 48h, hence, the performance was better than using the Adaptive PoPA alone. However, the uncertainty (previously unknown until now, but expected to be a normal Gaussian distribution) was essentially non-Gaussian (bimodal). Thus, further investigation as illustrated in Figure 12(a) and 12(b) shows that the Kalman+Adaptive PoPA performs better as the variance of forecast error is reduced when the uncertainty is normally distributed. Figure 12(b) shows the converter logic. Hence, a more sophisticated approach when the uncertainty is unknown should suffice.

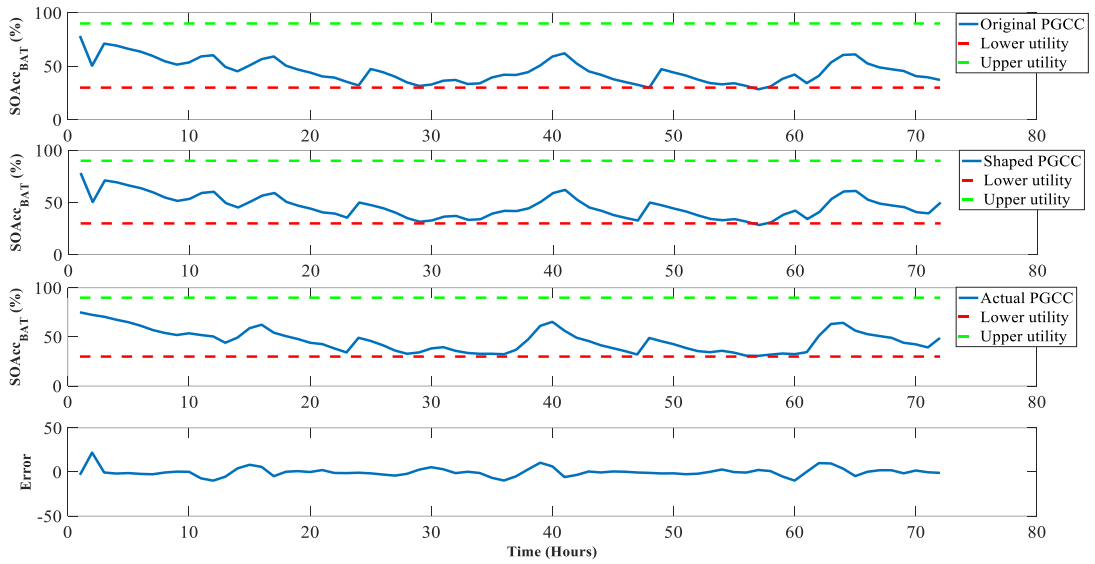


a)

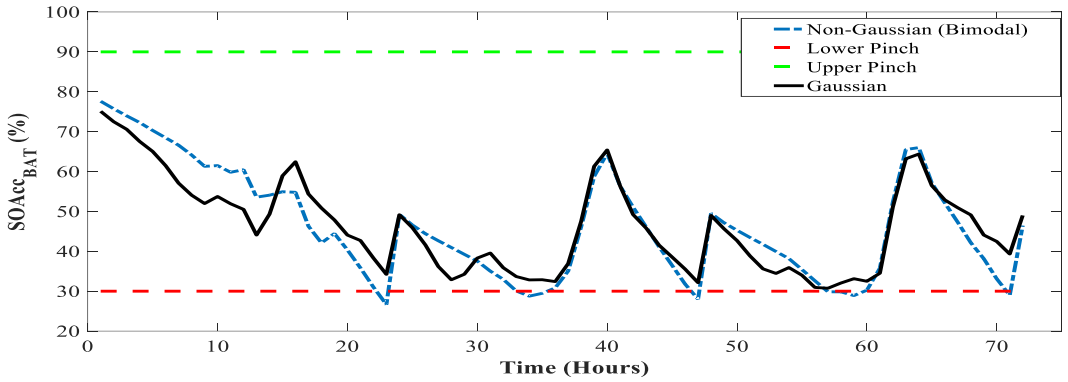


b)

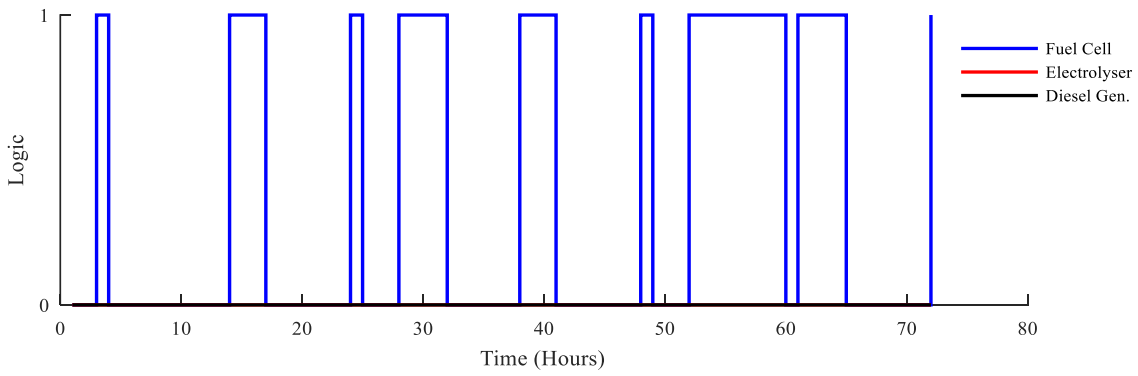
Fig 11. (a) The estimated and real Battery $SOAcc$ response with the Kalman Adaptive PoPA for 72 h under Gaussian uncertainty; (b) converter logic



(a)



b)

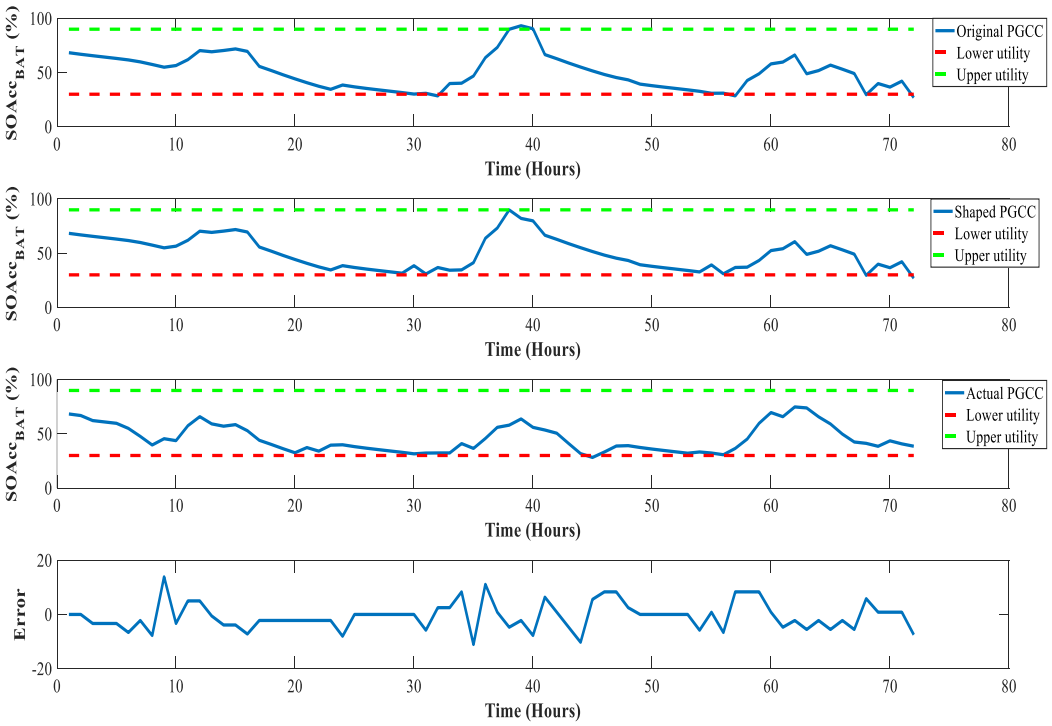


c)

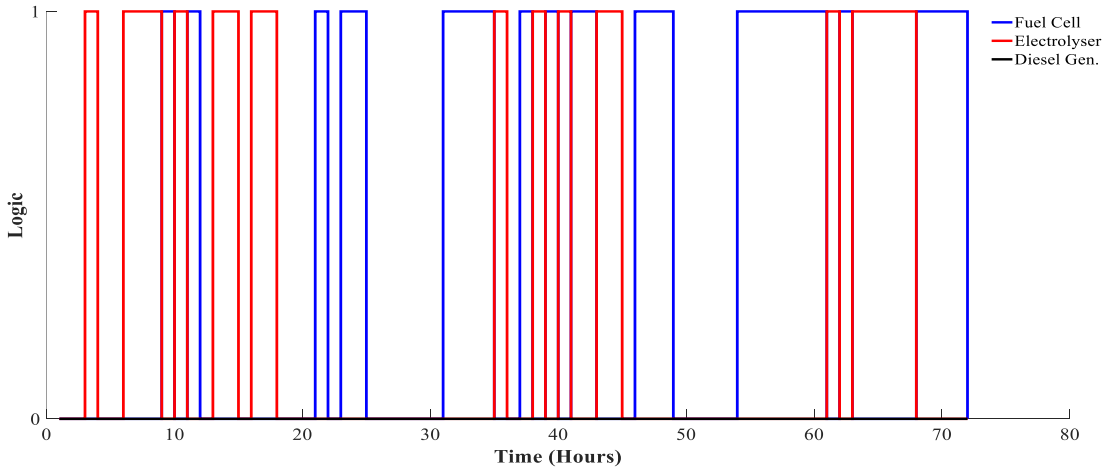
Fig 12: a) The estimated and real Battery $SOAcc$ response with the Kalman Adaptive PoPA for 72 h under Non-Gaussian (Bimodal) uncertainty, b) Comparison of the real $SOAcc$ response under both Gaussian and Non-Gaussian uncertainty, and c) converter logic under non-Gaussian uncertainty.

6.1.4 RL+Adaptive PoPA

The RL+Adaptive PoPA had only one violation of S_{LO}, which occurred at the 45th h as shown in Figure 13a. Also, the DSL was never activated. However, the FC and EL were activated 28 and 20 times respectively in a bid to track the Adaptive PoPA's PGCC as shown in Figure 13b.



a)



b)

Fig. 13. (a) shows the performance of the RL+Adaptive Pinch strategy for 72h; (b) converter logic

The violation of S_{LO} as indicated in Table 2, evidently showed Kalman Adaptive PoPA had the most significant improvement from 7 to 0 S_{LO} violations and none for the S_{UP} under Gaussian uncertainty and non-Gaussian case respectively. The RL Adaptive had no limit violations under the Gaussian uncertainty. While the Adaptive PoPA had an improvement when the uncertainty was Gaussian, there was negligible in the DA-PoPA's performance.

Table 2

Summary of the performance indices for 72h.

	Non-Gaussian Uncertainty				Gaussian Uncertainty			
	DA-PoPA	Adaptive PoPA	Kalman+ Adaptive PoPA	RL+ Adaptive PoPA	DA-PoPA	Adaptive PoPA	Kalman + Adaptive PoPA	RL + Adaptive PoPA
Lower Pinch violation	14	7	7	1	13	3	0	0
Upper Pinch violation	0	0	0	0	0	0	0	0
DSL Activation	2	0	0	0	4	0	0	0

6.2 Long-term operation

The proposed methods are evaluated against the DA-PoPA over a period of 8760 h and the results are shown in Table 3. From Table 3, the DA PoPA method had the worst performance indices as regards excessive charging of BAT ($SOAcc^n_{BAT}$>90%) and over-discharging ($SOAcc^n_{BAT}$<30%) and consequently fossil fuel utilisation due to the DSL activation, despite a decently sized HT of 15m^3 (initialised with $SOAcc^n_{HT}$ at 100%). The lower limit ($SOAcc^n_{BAT}$<30%) of the BAT was violated 804 times and accordingly the DSL was activated 229 times. Also the upper pinch limit ($SOAcc^n_{BAT} > 90\%$) of the BAT was violated 756 times.

Thus, benchmarked against the performance of the DA, the Adaptive, Kalman+Adaptive and RL+Adaptive PoPA methods led to a reduction in S_{LO} violation by 66%, 92% and 94%, as well as a decrease in the upper limit violation by 60%, 65% and 70%, respectively. Additionally, the DSL was activated only once with the Adaptive PoPA and was never activated with the Kalman, and RL+Adaptive PoPA. Consequently, a reduction in fossil fuel emission by 99.59%, 100% and 100% was achieved with the Adaptive, Kalman, RL+Adaptive PoPA EMS respectively. Furthermore, the reduction in upper limit violation by the Adaptive, Kalman and RL+Adaptive PoPA methods led to an increase in PV penetration by 6%, 6% and 7% respectively, due to the decreased violation of the PV (ON/OFF) protection constraint.

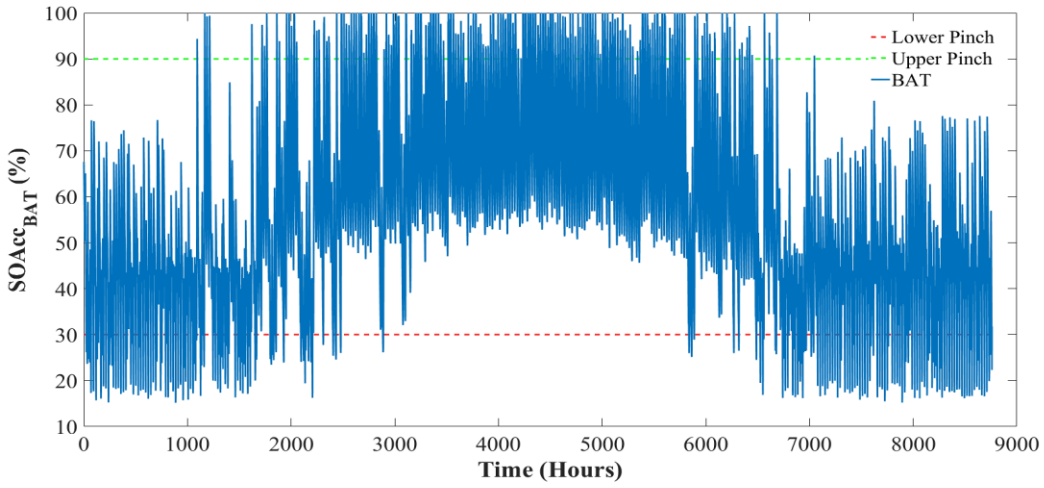
The RL+Adaptive method had the best performance with the least violations of S_{LO} and S_{UP}. However, to counteract the uncertainty, the learning agent increased activation of the FC and EL in the control horizon by 642% and 425% respectively, compared to the dictate of the Adaptive PoPA in the predictive horizon.

Also, the activation of the FC and EL with the Adaptive PoPA was seen to have increased by 95% and 150% and similarly for the Kalman +Adaptive PoPA, it was 520% and 255 % respectively, compared to the DA-PoPA.

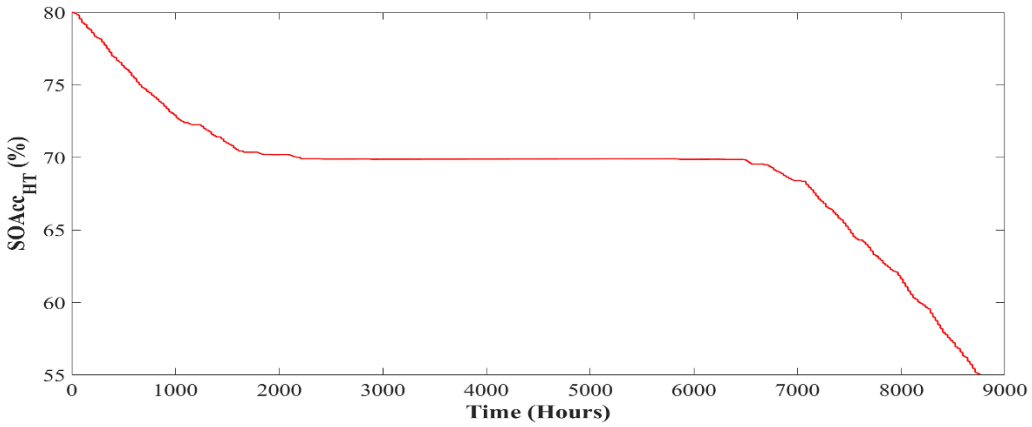
The available hydrogen in HT at 8760 hrs is as follows: 55% (DA-PoPA), 45% (Adaptive), 44% (RL+Adaptive) and 19% (Kalman+Adaptive). The $SOAcc_{HT}^{n}$ and $SOAcc_{BAT}^{n}$ are shown in Figure 14-17. The Kalman+Adaptive PoPA had the most usage of the hydrogen energy carrier, with the DA-PoPA having the least utilisation.

Table 3
Performance metrics characterizing the proposed Pinch methods for one year (8760 hr) with HT Volume of 15m^3.

	Day – Ahead PoPA	Adaptive PoPA	Kalman+Adaptive PoPA	RL+Adaptive PoPA
Lower Pinch violation ($SOAcc_{BAT}^{n}$ < 30%)	804	271	64	51
Upper Pinch violation ($SOAcc_{BAT}^{n}$>90%)	756	303	265	226
FC start-stop (cycles/year)	296	577	1837	3802
EL start-stop (cycles/year)	262	654	931	3503
DSL start-stop (cycles/year)	229	1	0	0
PV start-stop (cycles/year)	8004	8457	8495	8534

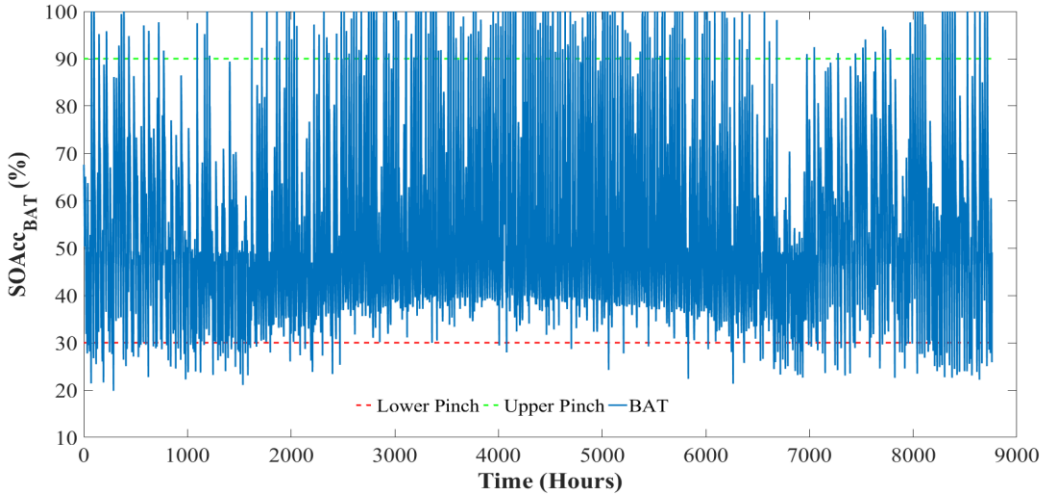


(a)

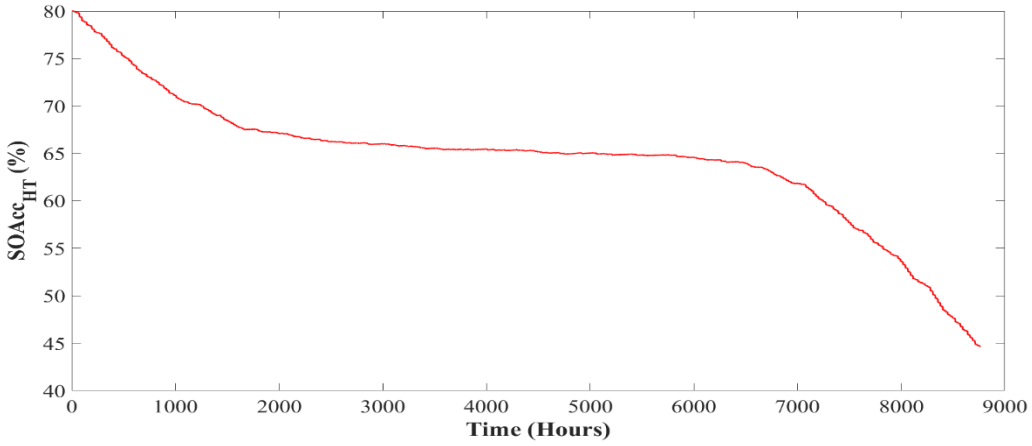


(b)

Fig.14. (a) The response of the BAT and (b) HT with the DA-PoPA

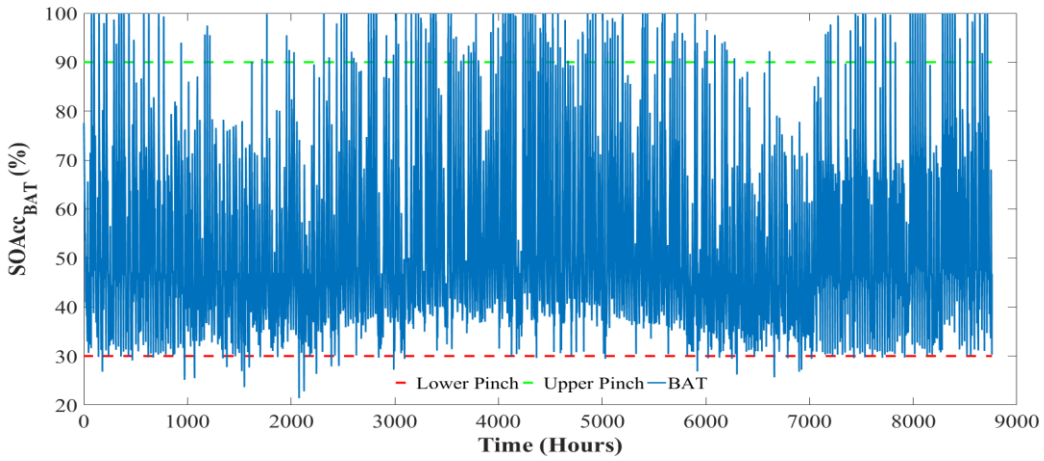


(a)

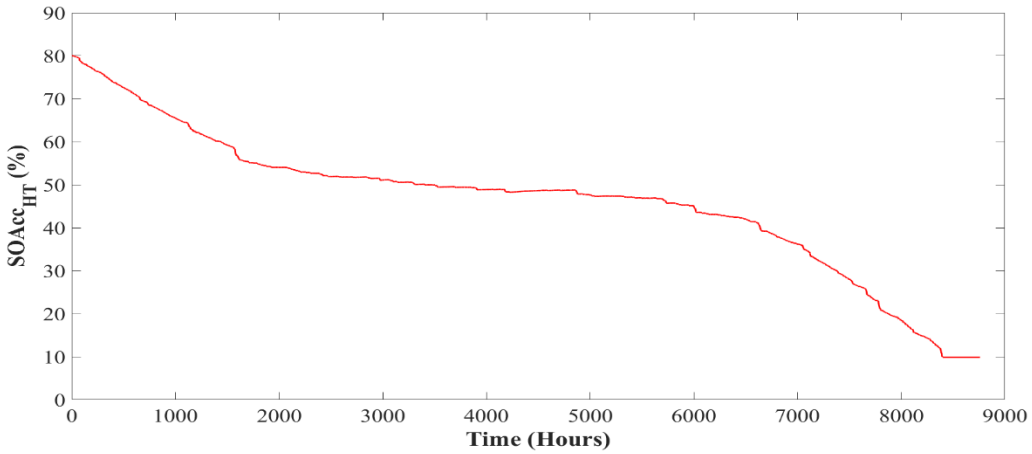


(b)

Fig. 15. (a) The response of the BAT and (b) HT with the Adaptive PoPA method

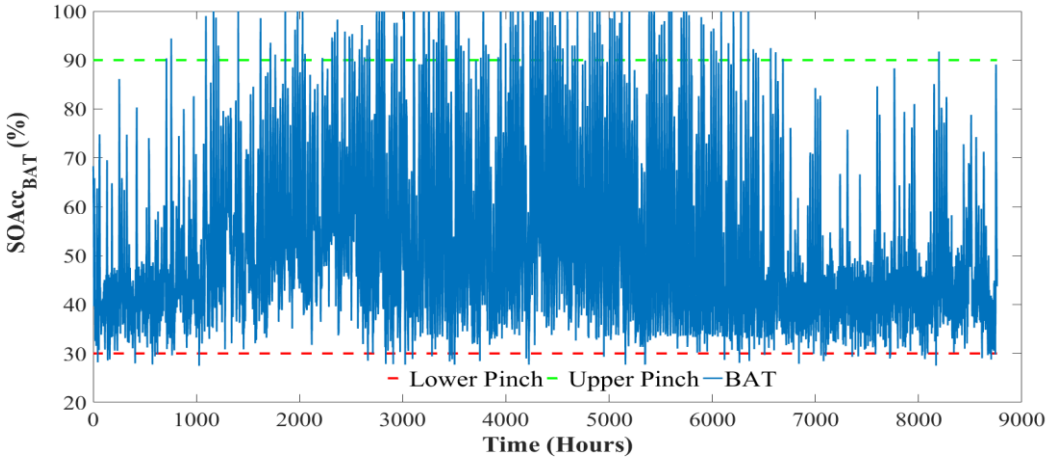


(a)

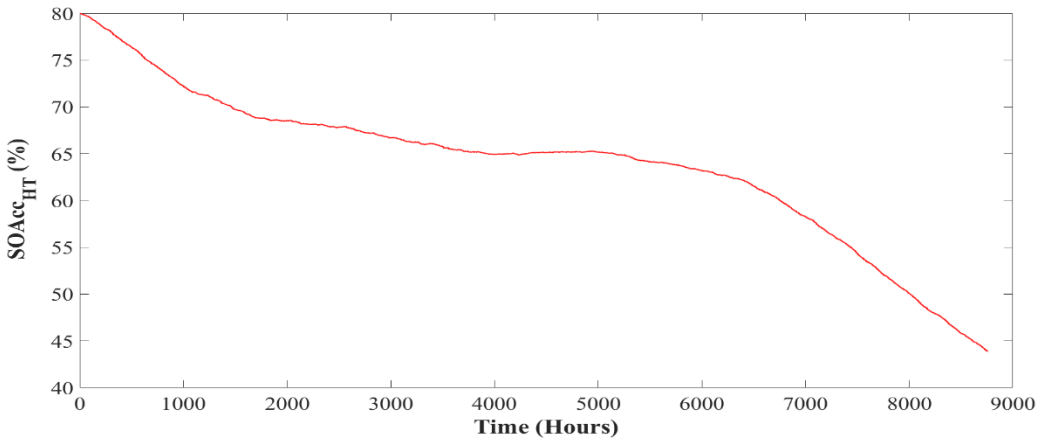


(b)

Fig. 16. (a) The response of the BAT and (b) HT response with Kalman +Adaptive PoPA



(a)



(b)

Fig. 17. (a) The response of the BAT and (b) HT response using RL+Adaptive Pinch Analysis

6.3 Sensitivity Analysis of HT Size with the PoPA Schemes

As shown in Figure 18, a sensitivity analysis was carried out to investigate the impact of hydrogen uncertainty by varying the HT capacity between 10, 5, and 1 m^3 with the EMS's. The RL+Adaptive PoPA scheme with HT at 10 m^3 had the fewest S_{LO} and S_{UP} violations of 68 and 256 times respectively, with the DSL never activated. The Kalman Adaptive PoPA had an S_{LO} and S_{UP} violation of 264 and 87 times. The DA-PoPA S_{LO} and S_{UP} violations were 756 and 804 times, and the adaptive PoPA violations were 303 and 271. However, the Kalman Adaptive PoPA activated the DSL at 15 instances in response to 87 lower limit violations, compared to the Adaptive PoPA which activated the DSL only once. Decreasing the HT capacity to 5 m^3 and 1 m^3, the RL+Adaptive PoPA lower limit was violated 1553 and 2616 times respectively, which consequently lead to the activation of the DSL 440 and 782 times.

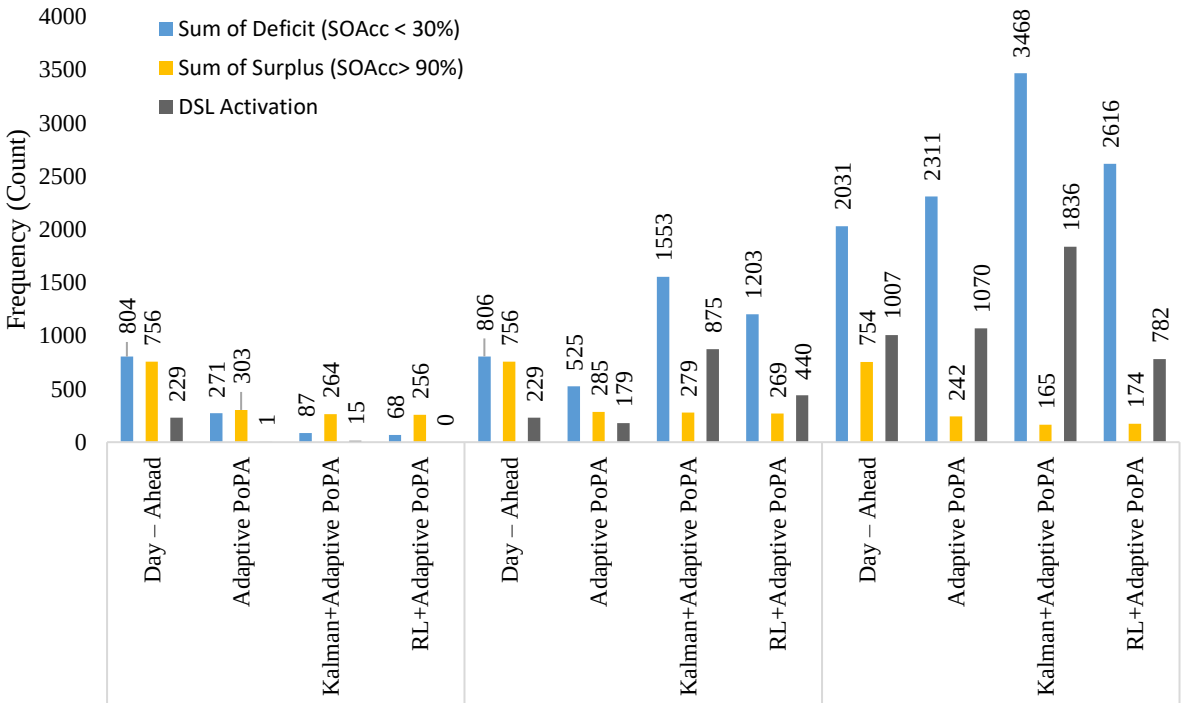


Fig.18. Sensitivity analysis of the PoPA Energy Management Schemes with 10, 5 and 1m^3 HT capacity.

When considering upper limit violations for different HT sizes, the RL+Adaptive PoPA had the best upper limit violation for an HT of 10m^3 and 5 m^3, and the second-best upper limit violation with an HT of 1m^3.

The RL+Adaptive PoPA had the least DSL activation overall for HT sizes of 5m^3 and 1m^3, which consequently implies that despite the S_{LO} violation of 1203 and 2616 times in that order were only better than the Kalman Adaptive PoPA's 1553 and 3468 times respectively. In addition, as seen in Figure 17, the preventive methods were more effective when the hydrogen is adequately available (i.e. HT > 5 m^3) (see figure I in appendix III).

The DA-PoPA violation of the upper limit remained almost unchanged despite the HT size variation. This clearly indicates the weakness of the DA-PoPA to uncertainty, in event of an unanticipated excess or deficit energy not considered during the daily energy target planning.

7. Conclusion

The Adaptive, Kalman+Adaptive and RL+Adaptive PoPA methods have been proposed to counteract uncertainty caused by PV and load profile variation which may impact the reliability of the HESS. These methods were compared against the existing DA-PoPA strategy using real-world data. The Adaptive PoPA had a better performance than the DA-PoPA, as a result of the inclusion of a feedback loop which minimised the effect of forecast deviations. However, the method offered a reactive strategy whose correction mechanism relied on the occurrence of the forecast error. Furthermore, the Adaptive PoPA incorporated a receding horizon without uncertainty propagation. The Kalman + Adaptive PoPA had a better performance than the adaptive PoPA. However, the formulation of the estimator relies on the assumption of a normally distributed uncertainty which was not the case. The RL+Adaptive method, which incorporates a learning agent illustrated for short and long-term operation, was shown to maximise the expected reward by acting optimally to meet the identified pinch targets. The RL+Adaptive had the best response across all performance indices; S_{LO} and S_{UP} limits violation as

well as reduced diesel carbon footprint when the HT was sized at 10m^3. However, even though the RL +Adaptive PoPA method offers the best results with respect to an avoided violation of operating limits on the storage devices this excellent performance comes at the cost of increased complexity. Therefore, the method used will be dependent on the application. For example, if there is a high confidence in the load/weather forecast then the DA PoPA method can be used, but if there is some error in the forecast, then the first Adaptive PoPA method, which does not require heavy processing power but is less accurate, should be used. However, if the difference between the real and the forecasted load/weather profile is significant and the uncertainty has specific statistical properties, then the right choice should be the use of the Adaptive PoPA with Kalman filter. Finally, if the error is large with no information about the type of uncertainty, then the RL+Adaptive PoPA should be the choice.

Appendixes

A.1 Pseudo Codes for the Proposed Algorithms

a. Pseudo Code for the First Proposition

1. Define the entire time span and intervals.
2. Define the initial systems state and EMS propositions
3. For all intervals k
   Perform within the prediction horizon the following procedures:
4. if $(k - N) = 23 \vee \Delta H(k) > (\xi == 5\%)$
   - 4.1.1 Repeat while Loop, L < =24 $\wedge$ ($S_{max} > S_{Up}^{l} \vee S_{min} < S_{Lo}^{l}$)
   - 4.2 Compute the PGCC with dispatch control sequence U_c according to equations (1)
   - 4.3 Determine $S_{min} = \underset{k\in[k,k+1,\dots,N]}{min}\ SOAcc_l^m(k)$ and $S_{max} = \underset{k\in[k,k+1,\dots,N]}{max}\ SOAcc_l^m(k)$
     - 4.3.1 If $S_{min} < S_{Lo}^{l}$
       a. Determine the energy $MOES = Lo - S_{min}$ required to shift the PGCC (Such that, $SOAcc_l^{m,1}(k_1) = (SOAcc_l^{m,0}(k_1) + MOES) < S_{Up}^{l}$)
       b. $U_c = FC\ :\ U_c(SOAcc_l^m) = [U_k(S_{k+1}), \dots U_{N-1}(S_T), \mid S_{k+1\,:\,k\in[1,2,\dots,N]} < S_{Up}^{l}]$ In a memory location, store the control sequence U_c
       c. Activate the selected converter U_c to inject the energy determined in step 4.2.1(a) at k_1 then go to step 4.3.
     - 4.3.2 if $S_{max} > S_{Up}^{l}$
       a. Determine the amount of energy MEES = $S_{max} - S_{Up}^{l}$ (Such that, $SOAcc_l^{m,1}(k_1) = (SOAcc_l^{m,0}(k_1) - \text{MEES}) > S_{Lo}^{l}$ to shift the PGCC).
       b. Activate the selected converter U_c , $c \in \{\text{EL}\}$ to absorb the energy determined in step 4.2.2(a) at k_1 then go to step 4.3.
   - 4.4 Determine $SOAcc_l^{m,L}(N)$: L ∈ [0: 24]
     - 4.4.1 if $SOAcc_l^{m,L}(N-1) \cong SOAcc_l^{m,L}(k_1)$
       a. calculate $\Delta S = SOAcc_l^{m,L}(k_1) - SOAcc_l^{m,L}(N-1)$ (such that $SOAcc_l^{m,1}(N-1)$ $= SOAcc_l^{m,0}(N-1) \pm \Delta S$
       b. Activate the selected converter U_c to inject or absorb the energy $\pm\,\Delta S$ determined in step 4.3.1(a) at $N-1$.
       c. repeat from step 4 until L>24
5. Activate the determined control sequence in control horizon $U_c(SOAcc_l^n) : S_{Lo}^{l} < [U_k(S_{k+1}), \dots U_{N-1}(S_N), \mid S_{k+1\,:\,k\in[1,2,\dots,N]} < S_{Up}^{l}]$
6. Determine state estimation error due to uncertainty:
   $\Delta H(k) = |SOAcc_{BAT}^{n}(k|k-1) - SOAcc_{BAT}^{m}(k)|$
7. Update the model with the actual system state with (7) for new PGCC re-computation
8. Repeat from step 3 until k > 8760

b. Pseudo Code for the Second Proposition

This follows steps 1 – 5 of the first proposal, but with the inclusion of the Kalman filter.

7. Update the priori covariance estimate $\mathcal{P}_k = \left[\mathcal{I} - \mathcal{K}_{\mathcal{G}}\mathcal{I}\right] \mathcal{P}_{k-1}$
8. Determine the Kalman gain $K_{G(k)} = \mathcal{P}_k \, I^T \, [\mathcal{I} \, \mathcal{P}_k \, \mathcal{I}^T + \, \mathcal{R}_k]^{-1}$
9. Predict the system state with the most recent output measurement from (11):

$$SOAcc_l^m \, (k) = SOAcc_l^m \, (k|k-1) + \mathcal{K}_{\mathcal{G}}(SOAcc_l^n \, (k) - \mathcal{I} \, SOAcc_l^m \, (k|k-1))$$

10. Estimate the posterior covariance matrix $\mathcal{P}_{k+1} = A \, \mathcal{P}_k A^T + \mathcal{R}_k$
11. Repeat from step 3 while $k \leq \, 8760$

c. Pseudo Code for the Third Proposition

This follows steps 1 – 6 of the first proposal, with the inclusion of the Q-learning state-action pair $Q(s, a)$.

5. Observe the systems state, s
6. For $k\sim = N$

   Switch ON/OFF dispatchable energy resources with the action selection policy $\pi \, (s)$ defined in (17) based on the state-action value function $Q(s, a)$.

   Else

   Override the action selected from policy $\pi \, (s)$ with AEEND EMS from Adaptive PoPA

   End
7. Observe $SOAcc_{BAT}^n$and determine the reward, R according to (21)
8. Update $Q(s, a)$ based on equation (16)

   s← s'
9. Randomly draw without replacement *n*-sample from memory $D \in <$ S, A, R, S', A' $>$ pairs of the most recent n-pinch limits violation experience due to uncertainty.
10. Update $Q(s, a)$ with the uncertainty experience
11. Repeat from step 3 until $k > 8760$

Table A.1

Connection	**Symbol**	**Logic proposition for HESS**
$BAT \leftarrow PV$	$\varepsilon_{PV}(t)$	$\bigcap_c [\varepsilon_{PV}^c(\mathrm{t})], \in \{Avl, Req, Gen\}$
	$\varepsilon_{PV}^{Avl}(t)$	1
	$\varepsilon_{PV}^{Req}(t)$	$q_{PV}^{SOAcc\ BAT}(\mathrm{t})$
	$\varepsilon_{PV}^{Gen}(t)$	1
	$q_{PV}^{SOAcc\ BAT}(\mathrm{t})$	$SOAcc_{\ BAT}(t) < S_{LO}^{BAT \leftarrow PV}(t)$
$BAT \leftarrow DSL$	$\varepsilon_{DSL}(t)$	$\bigcap_c [\varepsilon_{DSL}^c(\mathrm{t})], \in \{Avl, Req, Gen\}$
	$\varepsilon_{DSL}^{Avl}(t)$	1
	$\varepsilon_{DSL}^{Req}(t)$	$q_{DSL}^{SOAcc\ BAT}(\mathrm{t})$
	$\varepsilon_{DSL}^{Gen}(t)$	1
	$q_{DSL}^{SOAcc\ BAT}(\mathrm{t})$	$SOAcc_{\ BAT}(t) < S_{LO}^{BAT \leftarrow DSL}(t) \vee$ $\left[\begin{matrix}[S_{LO}^{BAT \leftarrow DSL}(t) < SOAcc_{\ BAT}(t) < S_{UP}^{BAT \leftarrow DSL}(t)] \wedge \\ [\ \varepsilon_{DSL}(t-1)]\end{matrix}\right]$
$BAT \leftarrow FC$	$\varepsilon_{FC}(t)$	$\bigcup_c [\varepsilon_{FC}^c(\mathrm{t})] \wedge \varepsilon_{FC}^{Avl}(t), \ c \in \{Req, Gen\}$
	$\varepsilon_{FC}^{Avl}(t)$	$\bigcap_l [a_{FC}^{SOAcc\ l}(t)], l \in \{HT, WT\}$
	$\varepsilon_{FC}^{Req}(t)$	$q_{FC}^{SOAcc\ BAT}(\mathrm{t})$
	$\varepsilon_{FC}^{Gen}(t)$	$\rho_{FC}^{U_c}$

Cont. of Table A.1

	$\rho_{FC}^{U_c}$	$\underset{\rightarrow}{U_c}(SOAcc_{BAT}(t))$
	$q_{FC}^{SOAcc\ BAT}(t)$	$SOAcc_{BAT}(t) < S_{LO}^{BAT \leftarrow FC}(t)$
	$a_{FC}^{SOAcc\ WT}(t)$	$SOAcc_{WT}(t) < S_{UP}^{WT \leftarrow FC}(t)$
	$a_{FC}^{SOAcc\ FT}(t)$	$SOAcc_{FT}(t) > S_{LO}^{FC \leftarrow HT}(t)$
$BAT \rightarrow EL$	$\varepsilon_{EL}(t)$	$\bigcup_{c}[\varepsilon_{EL}^{c}(t)] \cap \varepsilon_{PV}^{Avl}(t),\ c \in \{Req, Gen\}$
	$\varepsilon_{EL}^{Avl}(t)$	$\bigcap_{l}[a_{EL}^{SOAcc\ l}(t)], l \in \{BAT, HT\}$
	$\varepsilon_{EL}^{Req}(t)$	$q_{EL}^{SOAcc\ BAT}(t)$
	$\varepsilon_{EL}^{Gen}(t)$	$\rho_{EL}^{U_c}$
	$a_{EL}^{SOAcc\ BAT}(t)$	$SOAcc_{BAT}(t) > S_{LO}^{BAT \rightarrow EL}(t)$
	$a_{EL}^{SOAcc\ FT}(t)$	$SOAcc_{FT}(t) < S_{UP}^{EL \rightarrow HT}(t)$
	$q_{EL}^{SOAcc\ WT}(t)$	$SOAcc_{WT}(t) > S_{LO}^{EL \leftarrow WT}(t)$
	$\rho_{EL}^{U_c}$	$\underset{\rightarrow}{U_c}(SOAcc_{BAT}(t))$

Fig. A.1

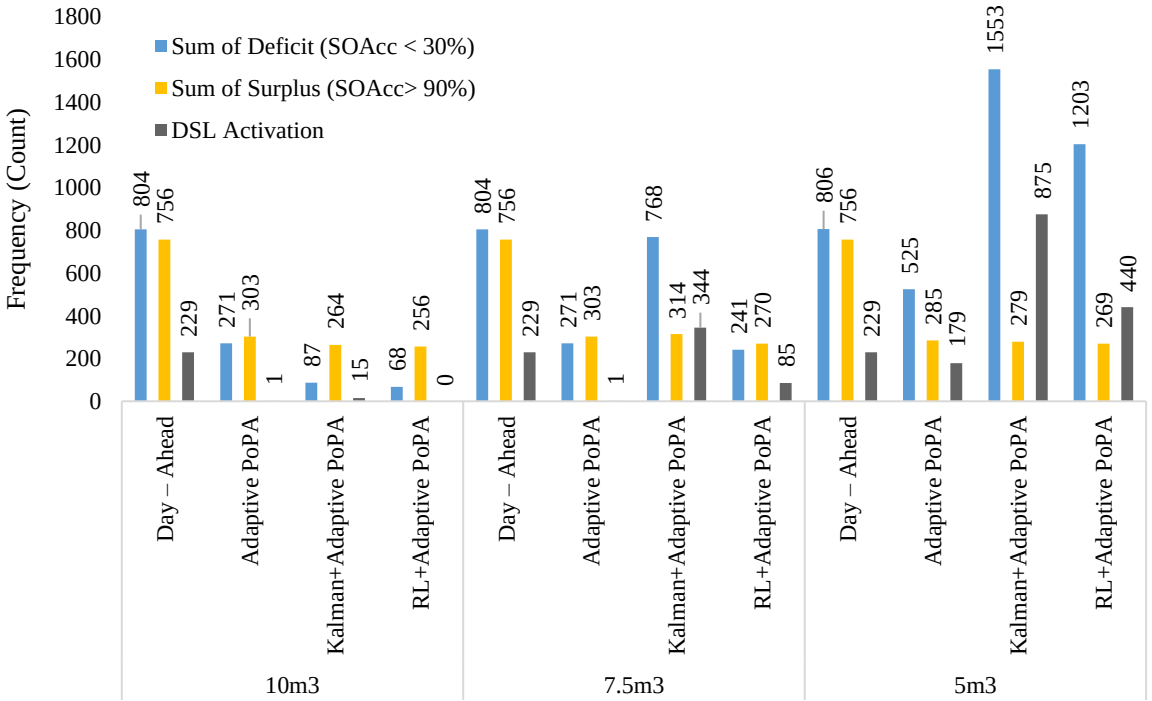


Fig. A.1. Sensitivity analysis of the PoPA Energy Management Schemes with 10, 7.5 and 5m^3 HT capacity.

References

[1] Qiao, L. A summary of optimal methods for the planning of stand-alone microgrid system. Energy and Power Engineering, 2013; 5(04), pp. 992.

[2] Muljadi, E., Wang, C. and Nehrir, M.H. June. Parallel operation of wind turbine, fuel cell, and diesel generation sources. In Power Engineering Society General Meeting, 2004; IEEE (pp. 1927-1932). IEEE.

[3] Dorji, T., Urmee, T. and Jennings, P., Options for off-grid electrification in the Kingdom of Bhutan. Renewable Energy, 2012; 45, pp.51-58.

[4] Kellogg, W.D., Nehrir, M.H., Venkataramanan, G. and Gerez, V. Generation unit sizing and cost analysis for stand-alone wind, photovoltaic, and hybrid wind/PV systems. IEEE Transactions on energy conversion, 1998; 13(1), pp.70-75.

[5] Giaouris, D., Papadopoulos, A.I., Voutetakis, S., Papadopoulou, S. and Seferlis, P. A power grand composite curves approach for analysis and adaptive operation of renewable energy smart grids, Clean Technologies and Environmental Policy, 2015; 17(5), pp.1171-1193.

[6] Bocklisch, T. Hybrid energy storage systems for renewable energy applications. Energy Procedia, 2015; 73, pp.103-111.

[7] Giaouris, D., Papadopoulos, A.I., Ziogou, C., Ipsakis, D., Voutetakis, S., Papadopoulou, S., Seferlis, P., Stergiopoulos, F. and Elmasides, C. Performance investigation of a hybrid renewable power generation and storage system using systemic power management models. Energy, 2013; 61, pp.621-635.

[8] http://www.systems-sunlight.com/, [Accessed 1st Nov 2017].

[9] Ipsakis, D., Voutetakis, S., Seferlis, P., Stergiopoulos, F., & Elmasides, C. Power management strategies for a stand-alone power system using renewable energy sources and hydrogen storage. International journal of hydrogen energy, 2009; 34(16), 7081-7095.

[10] Mahmood, H., Michaelson, D. and Jiang, J. A power management strategy for PV/battery hybrid systems in islanded microgrids. IEEE Journal of Emerging and Selected topics in Power electronics, 2014; 2(4), pp.870-882.

[11] Zhao, H., Wu, Q., Hu, S., Xu, H. and Rasmussen, C.N. Review of energy storage system for wind power integration support. Applied Energy, 2015; 137, pp.545-553.

[12] Fragiacomo, P., De Lorenzo, G. and Corigliano, O. Performance Analysis of an Intermediate Temperature Solid Oxide Electrolyzer Test Bench under a CO2-H2O Feed Stream. Energies, 2018; 11(9), p.2276.

[13] Mougin, J., Mansuy, A., Chatroux, A., Gousseau, G., Petitjean, M., Reytier, M. and Mauvy, F. Enhanced performance and durability of a high temperature steam electrolysis stack. Fuel Cells, 2013; 13(4), pp.623-630

[14] Jiao, K., He, P., Du, Q. and Yin, Y. Three-dimensional multiphase modeling of alkaline anion exchange membrane fuel cell. International Journal of Hydrogen Energy, 2014; 39(11), pp.5981-5995.

[15] Jiao, K., Huo, S., Zu, M., Jiao, D., Chen, J. and Du, Q. An analytical model for hydrogen alkaline anion exchange membrane fuel cell. International Journal of Hydrogen Energy, 2015; 40(8), pp.3300-3312.

[16] Olatomiwa, L., Mekhilef, S., Ismail, M.S. and Moghavvemi, M. Energy management strategies in hybrid renewable energy systems: A review. Renewable and Sustainable Energy Reviews, 2016; 62, pp.821-835.

[17] Wieczorek, M. and Lewandowski, M. A mathematical representation of an energy management strategy for hybrid energy storage system in electric vehicle and real time optimization using a genetic algorithm. Applied energy, 2017; 192, pp.222-233.

[18] Wang, S., Tang, Y., Shi, J., Gong, K., Liu, Y., Ren, L. and Li, J. Design and advanced control strategies of a hybrid energy storage system for the grid integration of wind power generations. IET Renewable Power Generation, 2014; 9(2), pp.89-98.

[19] Herath, A., Kodituwakku, S., Dasanayake, D., Binduhewa, P., Ekanayake, J., & Samarakoon, K., Comparison of Optimization-and Rule-Based EMS for Domestic PV-Battery Installation with Time-Varying Local SoC Limits. Journal of Electrical and Computer Engineering, 2019.

[20] De Souza, B. P., Zeni, V. S., Sica, E. T., Pica, C. Q., & Hernandes, M. V. Fuzzy Logic Energy Management System in Islanded Hybrid Energy Generation Microgrid. In 2018 IEEE Canadian Conference on Electrical & Computer Engineering (CCECE) (pp. 1-5). IEEE, 2018, May.

[21] Li, X., Xu, L., Hua, J., Lin, X., Li, J. and Ouyang, M. Power management strategy for vehicular-applied hybrid fuel cell/battery power system. Journal of Power Sources, 2009; 191(2), pp.542-549.

[22] Yu, Z., Zinger, D. and Bose, A. An innovative optimal power allocation strategy for fuel cell, battery and supercapacitor hybrid electric vehicle. Journal of Power Sources, 2011; 196(4), pp.2351-2359.

[23] De Lorenzo, G., Andaloro, L., Sergi, F., Napoli, G., Ferraro, M. and Antonucci, V. Numerical simulation model for the preliminary design of hybrid electric city bus power train with polymer electrolyte fuel cell. International Journal of Hydrogen Energy, 2014; 39(24), pp.12934-12947.

[24] Du, J., Zhang, X., Wang, T., Song, Z., Yang, X., Wang, H., & Wu, X., Battery degradation minimization oriented energy management strategy for plug-in hybrid electric bus with multi-energy storage system. Energy, 2018; 165, 153-163.

[25] Aktas, A., Erhan, K., Özdemir, S., & Özdemir, E., Dynamic energy management for photovoltaic power system including hybrid energy storage in smart grid applications. Energy, 2018; 162, 72-82.

[26] Zhao, L. U. O., Wei, G. U., Zhi, W. U., Zhihe, W. A. N. G., & Yiyuan, T. A. N. G., A robust optimization method for energy management of CCHP microgrid. Journal of Modern Power Systems and Clean Energy, 2018; 6(1), 132-144.

[27] Zhang, Y., Fu, L., Zhu, W., Bao, X., & Liu, C., Robust model predictive control for optimal energy management of island microgrids with uncertainties. Energy, 2018; 164, 1229-1241.

[28] Buhmann, J. M., Gronskiy, A. Y., Mihalák, M., Pröger, T., Šrámek, R., & Widmayer, P., Robust optimization in the presence of uncertainty: A generic approach. Journal of Computer and System Sciences, 2018; 94, 135-166.

[29] Hadayeghparast, S., Farsangi, A. S., & Shayanfar, H., Day-Ahead Stochastic Multi-Objective Economic/Emission Operational Scheduling of a Large Scale Virtual Power Plant. Energy, 2019

[30] Hu, M. C., Lu, S. Y., & Chen, Y. H., Stochastic programming and market equilibrium analysis of microgrids energy management systems. Energy, 2016; 113, 662-670.

[31] Tabar, V. S., Jirdehi, M. A., & Hemmati, R., Energy management in microgrid based on the multi objective stochastic programming incorporating portable renewable energy resource as demand response option. Energy, 2017; 118, 827-839.

[32] Hu, M. C., Lu, S. Y., & Chen, Y. H., Stochastic programming and market equilibrium analysis of microgrids energy management systems. Energy, 2016; 113, 662-670.

[33] Cai, M., Huang, G., Chen, J., Li, Y., & Fan, Y., A generalized fuzzy chance-constrained energy systems planning model for Guangzhou, China. Energy, 2018; 165, 191-204.

[34] Li, Y., Yang, Z., Li, G., Zhao, D., & Tian, W., Optimal scheduling of an isolated microgrid with battery storage considering load and renewable generation uncertainties. IEEE Transactions on Industrial Electronics, 2019; 66(2), 1565-1575.

[35] Huang, Y., Wang, L., Guo, W., Kang, Q., & Wu, Q. Chance constrained optimization in a home energy management system. IEEE Transactions on Smart Grid, 2018; 9(1), 252-260.

[36] Bruni, G., Cordiner, S., Mulone, V., Sinisi, V., & Spagnolo, F. Energy management in a domestic microgrid by means of model predictive controllers. Energy, 2016; 108, 119-131.

[37] Xiang, C., Ding, F., Wang, W., He, W., & Qi, Y. MPC-based energy management with adaptive Markov-chain prediction for a dual-mode hybrid electric vehicle. Science China Technological Sciences, 2017; 60(5), 737-748.

[38] Li, X., Han, L., Liu, H., Wang, W., & Xiang, C. Real-time Optimal Energy Management Strategy for a Dual-Mode Power-Split Hybrid Electric Vehicle Based on an Explicit Model Predictive Control Algorithm. Energy, 2019.

[39] Giaouris, D., Papadopoulos, A.I., Patsios, C., Walker, S., Ziogou, C., Taylor, P., Voutetakis, S., Papadopoulou, S. and Seferlis, P. A systems approach for management of microgrids considering multiple energy carriers, stochastic loads, forecasting and demand side response. Applied Energy, 2018; 226, pp.546-559.

[40] Papadopoulos, A.I., Giannakoudis, G., Seferlis, P. and Voutetakis, S. Efficient design under uncertainty of renewable power generation systems using partitioning and regression in the course of optimization. Industrial & Engineering Chemistry Research, 2012; 51(39), pp.12862-12876.

[41] Bandyopadhyay S. Design and optimization of isolated energy systems through pinch analysis. Asia Pac J Chem Eng, 2011; 6:518–526

[42] Alwi, S.R.W., Rozali, N.E.M., Abdul-Manan, Z. and Klemeš, J.J. A process integration targeting method for hybrid power systems. Energy, 2012; 44(1), pp.6-10.

[43] Linnhoff, B. and Flower, J.R. Synthesis of heat exchanger networks: I. Systematic generation of energy optimal networks. AIChE Journal, 1978; 24(4), pp.633-642.

[44] Smith, R. Chemical process: design and integration. John Wiley & Sons, 2005.

[45] Varbanov, P.S., Fodor, Z. and Klemeš, J.J. Total Site targeting with process specific minimum temperature difference (ΔTmin). Energy, 2012; 44(1), pp.20-28.

[46] Norbu, S. and Bandyopadhyay, S. Power Pinch Analysis for optimal sizing of renewable-based isolated system with uncertainties. Energy, 2017; 135, pp.466-475.

[47] Rozali, N.E.M., Alwi, S.R.W., Manan, Z.A., Klemeš, J.J. and Hassan, M.Y. Process integration techniques for optimal design of hybrid power systems. Applied Thermal Engineering, 2013, 61(1), pp.26-35.

[48] Esfahani, I.J., Lee, S. and Yoo, C. Extended-power pinch analysis (EPoPA) for integration of renewable energy systems with battery/hydrogen storages. Renewable Energy, 2015; 80, pp.1-14.

[49] Giaouris, D., Papadopoulos, A.I., Seferlis, P., Voutetakis, S. and Papadopoulou, S. Power grand composite curves shaping for adaptive energy management of hybrid microgrids. Renewable Energy, 2016; 95, pp.433-448.

[50] Brka, A., Al-Abdeli, Y.M. and Kothapalli, G. Predictive power management strategies for stand-alone hydrogen systems: Operational impact. International Journal of Hydrogen Energy, 2016; 41(16), pp.6685-6698.

[51] Dias, L.S. and Ierapetritou, M.G. Integration of scheduling and control under uncertainties: Review and challenges. Chemical Engineering Research and Design, 2016; 116, pp.98-113.

[52] Richards, A. and How, J. Robust model predictive control with imperfect information." In Proceedings of the 2005, American Control Conference IEEE, 2005; pp. 268-273.

[53] Bemporad, Alberto, Francesco Borrelli, and Manfred Morari. "Min-max control of constrained uncertain discrete-time linear systems." IEEE Transactions on automatic control 2003; 48, (9), 1600-1606.

[54] Siswantoro, J., Prabuwono, A.S., Abdullah, A. and Idrus, B. A linear model based on Kalman filter for improving neural network classification performance. Expert Systems with Applications, 2016; 49, pp.112-122.

[55] Takeda, H., Tamura, Y. and Sato, S. Using the ensemble Kalman filter for electricity load forecasting and analysis. Energy, 2016; 104 pp.184-198.

[56] Al-Hamadi, H.M. and Soliman, S.A. Short-term electric load forecasting based on Kalman filtering algorithm with moving window weather and load model. Electric power systems research, 2004; 68(1), pp.47-59.

[57] Sutton, R.S. Learning to predict by the methods of temporal differences. Machine learning, 1988; 3(1), pp.9-44.

[58] Geramifard, A., Walsh, T.J., Tellex, S., Chowdhary, G., Roy, N. and How, J.P. A tutorial on linear function approximators for dynamic programming and reinforcement learning. Foundations and Trends® in Machine Learning, 2013; 6(4), pp.375-451.

[59] Watkins, C.J. and Dayan, P. Q-learning. Machine learning, 1992; 8(3-4), pp.279-292.

[60] ALTUNTAŞ, N., Imal, E., Emanet, N. and Öztürk, C.N. Reinforcement learning-based mobile robot navigation. Turkish Journal of Electrical Engineering & Computer Sciences, 2016; 24(3), pp.1747-1767

[61] Zhang, Q., Li, M., Wang, X. and Zhang, Y. Reinforcement Learning in Robot Path Optimization. JSW, 2012, 7(3), pp.657-662.

[62] Sutton, R.S. Dyna, an integrated architecture for learning, planning, and reacting. ACM SIGART Bulletin, 1991, 2(4), pp.160-163.

[63] Ernst, D., Glavic, M., Capitanescu, F. and Wehenkel, L. Reinforcement learning versus model predictive control: a comparison on a power system problem. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2009; 39(2), pp.517-529.

[64] Peng, K. and Morrison, C. Model Predictive Prior Reinforcement Leaning for a Heat Pump Thermostat. In IEEE International Conference on Automatic Computing: Feedback Computing (Vol. 16), July 2016

[65] Kuznetsova, E., Li, Y. F., Ruiz, C., Zio, E., Ault, G., & Bell, K. Reinforcement learning for microgrid energy management. Energy, 2013; 59, 133-146.

[66] François-Lavet, V., Taralla, D., Ernst, D., & Fonteneau, R. Deep reinforcement learning solutions for energy microgrids management. In European Workshop on Reinforcement Learning (EWRL 2016), 2016.

[67] Kofinas, P., Dounis, A. I., & Vouros, G. A. Fuzzy Q-Learning for multi-agent decentralized energy management in microgrids. Applied energy, 2018; 219, 53-67.

[68] Liu, T., Wang, B., & Yang, C. Online Markov Chain-based energy management for a hybrid tracked vehicle with speedy Q-learning. Energy, 2018; 160, 544-555.

[69] Rocchetta, R., Bellani, L., Compare, M., Zio, E., & Patelli, E. A reinforcement learning framework for optimal operation and maintenance of power grids. Applied Energy, 2019 241, 291-301.

[70] Xiong, R., Cao, J., & Yu, Q. Reinforcement learning-based real-time power management for hybrid energy storage system in the plug-in hybrid electric vehicle. Applied Energy, 2018, 211, 538-548.

[71] Lin, X., Wang, Y., Bogdan, P., Chang, N., & Pedram, M. Reinforcement learning based power management for hybrid electric vehicles. In Proceedings of the 2014 IEEE/ACM International Conference on Computer-Aided Design (pp. 32-38). IEEE Press, 2014.

[72] Nyong-Bassey, B.E., Giaouris, D., Papadopoulos, A.I., Patsios, H., Papadopoulou, S., Voutetakis, S., Seferlis, P., Walker, S., Taylor, P. and Gadoue, S. Adaptive Power Pinch Analysis for Energy management of Hybrid Energy Storage Systems. In Circuits and Systems (ISCAS), 2018 IEEE International Symposium on (pp. 1-5). IEEE, 2018 May.

[73] Nyong-Bassey B.E., Giaouris, D., Patsios, H., Gadoue, A. I., Papadopoulou, Seferlis P., Voutetakis, S., Papadopoulos S. 'A Probabilistic Adaptive Model Predictive Power Pinch Analysis (PoPA) Energy Management Approach to Uncertainty'. 9th International Conference on Power Electronics, Machines and Drives (PEMD). Journal of Engineering. IET, 2018.

[74] Tijsma, A.D., Drugan, M.M. and Wiering, M.A. December. Comparing exploration strategies for Q-learning in random stochastic mazes. In Computational Intelligence (SSCI), 2016 IEEE Symposium Series on (pp. 1-8) IEEE, 2016.

[75] Carden, S.W. Convergence of a reinforcement learning algorithm in continuous domains (Doctoral dissertation, Clemson University), 2014.

[76] Campbell, J.S., Givigi, S.N. and Schwartz, H.M. Multiple model Q-learning for stochastic asynchronous rewards. Journal of Intelligent & Robotic Systems, 2016; 81, (3-4), pp.407-422.

[77] Tokic, M. September. Adaptive ε-greedy exploration in reinforcement learning based on value differences. In Annual Conference on Artificial Intelligence Springer, Berlin, Heidelberg, 2010; (pp. 203-210).

[78] Bellman, R. Dynamic programming. Courier Corporation, 2013.

[79] http://data.ukedc.rl.ac.uk/simplebrowse/edc/efficiency/residential/LoadProfile/data [Accessed 1st Nov. 2017].

[80] http://pvwatts.nrel.gov/pvwatts.php, [Accessed 1st Nov 2017].