Reinforcement Learning Based Recloser Control for Distribution Cables with Degraded Insulation Level

Qiushi Cui, *Member, IEEE,* Yousaf Hashmy, *Student Member, IEEE,* Yang Weng, *Member, IEEE,* and Michael Dyer, *Member, IEEE*

***Abstract*—Utilities continuously observe cable failures on aged cables that have an unknown degraded basic insulation level (BIL). One of the root causes is the transient overvoltage (TOV) associated with circuit breaker reclosing. To solve this problem, researchers propose a series of controlled switching methods, most of which belong to deterministic control. However, in power systems, especially in distribution networks, the switching transient is buffeted by stochasticity. Since it is hard to model transient overvoltage due to its complexity, we propose a model-free stochastic control method for reclosers under the existence of uncertainty and noise. Concretely, to capture high-dimensional dynamics patterns, we formulate the recloser control problem by incorporating the temporal sequence reward mechanism into a deep Q-network (DQN). Meanwhile, we embed our physical understanding of the problem into the action probability allocation and develop an infeasible-action-space-elimination algorithm. Through PSCAD simulation, we first reveal the impact of load types on cables' TOVs. Then, to reduce the training burden for the proposed reinforcement learning (RL) control method in different applications, we establish a post-learning knowledge transfer method. After the validation with our industrial partner, we exhibit several learning curves to show the enhanced performance. The learning efficiency is proved to be outstanding due to the proposed time sequence reward mechanism and infeasible action elimination method. Moreover, the results on knowledge transfer demonstrate the capability of method generalization. Finally, a comparison with conventional methods is conducted. It illustrates the proposed method is most effective in mitigating the TOV phenomenon among three methods.**

***Index Terms*—Transient overvoltage, cable failure, controlled switching, reinforcement learning, post-learning knowledge.**

I. Introduction

SWITCHING voltage surge or transient is the result of energization or de-energization of the transmission or distribution lines and large electrical apparatuses such as reactors and capacitor banks. These actions can occur in the system due to system configuration changes or faults. During these conditions, the inductive or capacitive loads release or absorb the energy suddenly and generate voltage or current transient. Consequently, voltage surges may occur and, therefore, jeopardize the equipment and personal safety. Specifically, the switching surges usually occur upon the energization of lines, cables, transformers, reactors, or capacitor banks [1].

For long high-voltage lines, they store a large amount of energy, which generates sufficient voltage transients in the systems [2]. For conductors in distribution systems, there are two cases. Underground line capacitance for power cables is far higher as compared to their overhead counterparts due to closeness of cables and proximity to earth. As a result, underground lines have 20-75 times [3] the line charging current[1]. Thus, cables can trap a high amount of charge. The trapped charge is a residual charge in the line or cable subsequent to de-energization. If the trapped charge is with the same polarity as the system voltage, switching overvoltage may be observed. Although most papers focus on the transient overvoltage in transmission lines, cable failures due to TOV are continuously reported by utility (see Fig. 1). In fact, a slow TOV whose duration is less than a cycle should not be a problem for the insulation as the cable BIL is much higher. However, most aged cables have unknown and degraded BIL, causing frequent cable failures in modern smart grids. Besides, most utilities probably do not reclose into faults on underground systems, as faults in underground systems are considered permanent. The purpose of reclosing is to allow temporary faults to be cleared, which is typical for an overhead system. Practically, the primary purpose of this paper is to investigate what damaging effects reclosing into underground faults may produce and provide arguments to change this practice. Therefore, we are motivated to investigate the effects of reclosing, primarily the resulting overvoltage phenomenon in distribution systems, for the practical consideration of eliminating the occurrence of cable failure.

Figure 1. Photos of the failed cables after 5-recloses: the faulted cable (left) and the adjacent unfaulted cable (right) that has similar damage. Cause of reclosing: switch hit by a vehicle. Source of photos: Salt River Project.

To achieve the above target, tests on a real feeder is an unviable solution since the customers downstream will go through a power outage. Therefore, computer simulation of the field tests is developed to study the transient electromagnetic phenomena. Real-time system parameters and measurements are required to prepare system models and perform an ex-

This work was supported in part by the National Science Foundation of the United States under Grant 1810537, in part by the Salt River Project on "An Investigation of the Effects of Reclosing on Distribution Systems with Under-ground Cables", and in part by the ARPA-E Project on "Sensor Enabled Modeling of Future Distribution Systems with Distributed Energy Resources".

Q. Cui, Y. Hashmy and Y. Weng are with the Department of Electrical and Computer Engineering, Arizona State University, Tempe, AZ, 85281, USA E-mail: {qiushi.cui,shashmy,yang.weng}@asu.edu.

M. Dyer is with Salt River Project. E-mail: michael.dyer@srpnet.com.

[1]Capacitance causes current to flow even when no load is connected to the cable. This is called line charging current.

act transient study [4], [5]. This is very useful to identify available voltage surge, determine the equipment insulation coordination, and select protective equipment operating characteristic [1], [6]. However, it is essential to consider the peak over-voltage discrepancy between the frequency-based simulation model results and real-time field measurements. [7] presents some cases with a good agreement between simulation results obtained with an Electromagnetic Transients Program (EMTP)-type program and either field measurements or transient network analyzer results. In [8], researchers solve the above issue by modeling corona, prestrike voltage, and frequency-based line parameters appropriately.

The above works on modeling have underpinned the development of the device-based and control-based overvoltage mitigation methods. The power industry has witnessed the evolution of surge arresters from the air gap and silicon carbide types to the metal oxide varistors (MOV). In extra-high voltage applications, MOV and breaker with closing resistors are two basic methods to restrict switching surges [9], [10]. In high voltage transmission systems, switching surges are destructive to electrical equipment, so surge arresters are typically installed near large transformers and on line terminals to suppress surges [11]. Whereas in medium and low voltage levels, as the penetration of distributed energy resources gets deeper, it is still not clear whether the arresters are a viable solution. One thing is clear: it is not economical to place surge arresters all over the distribution networks due to their vast reaches. Besides surge arresters, other devices used to limit switching overvoltage include pre-insertion resistors [12], [13], and magnetic voltage transformers [14].

In addition to the device-based method, controlled switching belongs to the second category of overvoltage mitigation methods. The core of controlled switching is statistical switching, where the worst-case scenarios are determined through several dimensions of overvoltage scenarios. Statistical switching is adopted for decades [15]–[21]. Investigated scenarios include switching speed [22], actual operating capacity [21], load and line length [23], etc. For example, the impacts of the switching speed of the disconnector are studied by statistical methods in [22]. [24] mitigates the transient overvoltage by controlling the voltage conditions preceding voltage breakdowns in the disconnector contact system. A developed version of the transmission line zero-crossing controlled switching relay is proposed in [25], considering the polarity of trapped charge. [26] proposes a method to determine the optimum closing point for CB contacts without imposing any limitation on line side fluctuations.

Unlike the conventional controlled switching methods that rely on deterministic control, this paper views controlled switching as a stochastic control. In a deterministic model, the future state is theoretically predictable. Thus, most researchers look into the statistical switching overvoltage distributions for different switching operations, and then design the control according to the observation. However, in power systems, especially in distribution networks, the switching transient is buffeted by stochasticity. We need a stochastic model to possess inherent randomness and uncertainty. Unfortunately, relatively little has been done to develop a stochastic control mechanism that views the complexity of the control task as a Markov decision process (MDP). Since it is hard to assume knowledge or cost function of the overvoltage dynamics, we want to combine the advantages of off-policy control and value function approximation. Meanwhile, given the high-dimensional dynamic complexity of power systems, a deep RL method is re-designed to improve the control performance. Therefore, after the validation with our industrial partner, we propose a recloser control method using deep Q-networks (DQNs). The main contribution of this article is summarized below.

- Conventional controlled switching methods do not involve observation uncertainty and noise that drives the evolution of the system; therefore, we formulate the recloser control problem by incorporating the temporal sequence reward mechanism into a DQN to mitigate reclosing TOV. Meanwhile, we invent an infeasible-action-space-elimination algorithm through time-variant probability allocation in DQNs.
- To overcome the training burden for the proposed RL control method in different applications, we develop a post-learning knowledge transfer method for recloser control to handle complex system operating conditions, save training time, improve the recloser performance, and reduce the required data volume.

The rest of this article is framed as follows: Section II provides the background of the reclosing impact on underground cables. The proposed recloser control method using RL is elaborated in Section III. Section IV shows the numerical results, followed by the discussions in Section VI and the conclusions in Section VII.

II. Reclosing Impact on Cables via PSCAD

As mentioned earlier, one of the reasons for the failure of cable is TOVs. TOV can arise from the supply or from switching inductive loads, harmonic currents, DC feedback, mutual inductance, high-frequency oscillations, large starting currents, and large fluctuating loads [27]. TOV or surges are temporary high magnitude voltage peaks for a short duration of time, e.x., lightning. Switching transients in electrical networks often occurs. Although the voltage magnitude is lower than the lightning surge, the frequency at which it occurs causes aging of cable insulation and eventually breaks down resulting in flashover. To observe the TOVs in computer programs, we utilize the 750 MCM-AL [28] cable, which is widely implemented in the State of Arizona and many others. In this section, we focus on the modeling of switching and power systems.

A. Switching Modeling

For the switching modeling, we use the statistical breakers in PSCAD to account for the physical metal contact and the issue of pole span. Pole span is the time span between the closing instant of the first and the last pole. The single-pole operation of three-phase breaker is applied to incorporate the angle difference in the operation of different poles because of the mechanical inconsistencies. The resulting TOVs upon 100

simulations of different sets of circuit breaker closing times with a standard deviation of 4 in the half interval are shown in Table I. This table brings some flavors on how the pole span contributes to the maximum TOVs. One can refer to Section IV-A for the system parameters.

Table I
COMPARISON UNDER THREE TYPES OF POLE SPANS.

Pole span (ms)	Highest TOV (pu)	Avg. TOV (pu)
0	1.55	1.55
0.24 [1]	1.58	1.52
3.7 [1]	1.58	1.51

When the switching occurs at other angles, different TOVs are obtained. Although we did not demonstrate all the cases with higher TOVs, it shows that the optimal controlled switching time is crucial to TOV mitigation under the current switching modeling. It is noteworthy that the limitation of the adopted switch modeling is imperfect, the details of which can be found in Section VI. Meanwhile, it is evident that overvoltages frequently occur on cables; therefore, it is imperative to provide a solution that lowers the probability of cable failure.

B. Power System Modeling

Firstly, two different line models, namely, distributed line model and frequency-dependent π model, are employed to capture different aspects of cable characteristics. Fig. 2 and Fig. 3 demonstrate the reclosing waveform plotted using PSCAD. At the end of the cable, a capacitor bank and transformers are connected to represent the reality, which explains the occurrence of the resonance effect during recloser dead time [25]. In the case of a distributed line model, a TOV of 1.51 pu is observed when switching at zero degree of the source voltage. However, a TOV of 1.55 pu is observed when a frequency-dependent π model is used. In the majority of the cases, TOVs are higher with a π model because resistance, inductance and capacitance of the line are considered together. Secondly, a detailed three-phase voltage source model is selected from the PSCAD library. The associated parameters, in particular the source impedance, are adopted from our industry partner's realistic distribution feeders.

III. REINFORCEMENT LEARNING BASED RECLOSER CONTROL METHOD

It is important to select an RL method that is suitable for the particular problem under study. In general, RL is classified into model-based (MB) and model-free (MF). In MB RL, we choose the classical World model as an example. Since it is MB, an environmental model is needed during learning. However, given the complexity of the TOV problem under study, it is hard to construct an internal model of the transitions and immediate outcomes for recloser control. That is why we did not move forward with MB RL. In MF RL algorithms, the agent relies on trial-and-error experience to reach the optimal policy. The typical methods include policy optimization and Q-learning. Under the policy optimization approach, we select the popular one – policy gradient (PG) as a comparison.

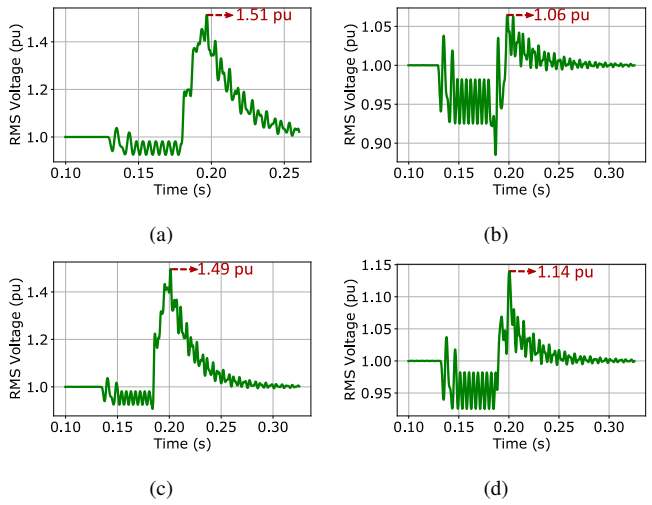

Figure 2. An example of the TOV waveform when the reclosing angles are set at 0°, 45°, 90° and 180°, respectively. The recloser opens at $t = 0.12$ sec, and closes at $t = 0.17$ sec. Tests are under lagging load condition, at a 12 kV feeder connecting with a 2.5 miles long cable using a distributed line model.

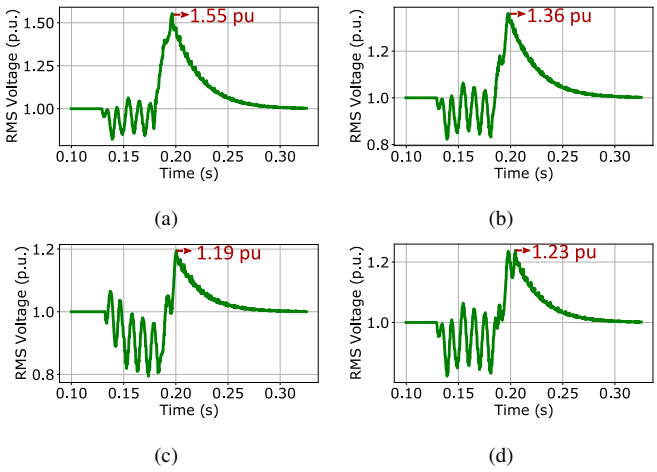

Figure 3. An example of the TOV waveform when the reclosing angles are set at 0°, 45°, 90° and 180°, respectively. The recloser opens at $t = 0.12$ sec, and closes at $t = 0.17$ sec. Tests are under lagging load condition, at a 12 kV feeder connecting with a 2.5 miles long cable using a frequency-dependent π model.

In contrast, for Q-learning methods, we choose the basic version and the advanced version DQN. Please note that this paper utilizes DQN method for RL control. The main advantage of DQN over PG in our case is that it involves discrete action space, while PG is for continuous action spaces. Our whole contention is to reduce the action space. However, PG method will consider 0, 1 and anything in between. Whereas, a breaker can have precisely two discrete actions (Off and On). Therefore, owing to the discrete nature of the action space involved in Q-learning-based RL, it is perhaps the best choice to reduce the computational burden. For the selection between Q-learning and DQN, we decide to go with DQN due to its powerful value function approximation capability in multiple power system scenarios. The above comparison of selecting the RL methods is summarized in Table II, where the circle highlights the main reason for this method not been selected.

In the remaining part of this section, we start with the

Table II
COMPARISON OF FOUR REINFORCEMENT LEARNING APPROACHES.

Item	World model [29]	Policy gradient	Q-learning	DQN
Model-based (MB) or model-free (MF)?	MB	MF	MF	MF
Need environmental model?	Yes	No	No	No
Based on value function?	No	No	Yes	Yes
Value function approximation?	No	No	No	Yes
Action Space	Continuous/ Discrete	Continuous	Discrete	Discrete

impetus of choosing the DQN algorithm, which is capable of dealing with the continuous status space of the recloser observation. To control the reclosers, we, next, elaborate on our design of temporal sequence reward mechanism, infeasible action space elimination algorithm, and the post-learning knowledge transfer method.

A. The Deep Q-network for Better Value Approximation

The task of TOV mitigation requires a model-free control algorithm that finds an optimal strategy for solving a dynamical control problem. Obviously, RL is a suitable solution. Among various types of RL algorithms, we believe the off-policy control where the agent usually uses a greedy policy to select actions can be incorporated with the action value estimation design. Therefore, we choose Q-learning to satisfy this requirement. Based on the complexity of the electric grids, we find that the value-based DQN method needs to involve intensive use of simulation for the parametric approximation. To enable self-learning of the recloser control, we adopt an actor-critic system to estimate the rewards. The *critic* in this system evaluates the value function, and the *actor* is the algorithm that improves the obtained value. DQN agents use the following training algorithm, in which they update their critic model at each time step. First, we need to initialize the critic $Q(s,a)$ with random parameter values θ_Q, and initialize the target critic with the target update smoothing method. Then, according to [30], at each time step:

1) With probability ε, select a random action A. Otherwise, select the action that maximizes the critic value function:
$$A = \arg\max_{A} Q(S,A|\theta_Q). \quad (1)$$
It makes sure the off-policy method always follows the greedy policy – the best action value estimations.
2) Execute action A, then calculate the reward R and the next state S'. If there are associated TOVs, they will be measured in this step, and the reward is calculated.
3) Store the experience (S,A,R,S') in the experience buffer. This technique smooths the training distribution over many past behaviors.
4) Randomly sample M experiences (S_i,A_i,R_i,S_i') from the experience buffer. We call the M sampled dataset the random mini-batch. If S_i' is a terminal state, set the value function target y_i to R_i. Otherwise set it to:
$$y_i = R_i + \gamma \max_{A'} Q'(S_i',A'|\theta_{Q'}), \quad (2)$$
where γ is the discount factor, and Q' is the value for the next state. In such a way, the current state that the recloser measures is represented in a form that the RL agent can interpret.
5) Update the critic parameters by one-step minimization of the loss L across all sampled experiences:
$$L = \frac{1}{M}\sum_{i=1}^{M}(y_i - Q(S_i,A_i|\theta_Q))^2. \quad (3)$$
Thereby, the parameter θ_Q for value approximation is calculated.
6) Update the target critic using the target smoothing update methods (τ is the smoothing factor):
$$\theta_{Q'} = \tau\theta_Q + (1-\tau)\theta_{Q'}. \quad (4)$$

B. Temporal Sequence Reward to Guarantee Learning Quality

To develop a DQN to mitigate TOVs, we first consider its state design. For each phase $p \in \{A,B,C\}$, there are voltage and current measurements from the bus located downstream of the breaker under study. Similar to conventional recloser, the magnitudes of voltage $|V_p|$ and current $|I_p|$ along with the voltage phase angle θ_{V_p} and current phase angle θ_{I_p} of the measurements are selected for defining a 4-dimensional state space $\mathbf{s}$ of the system:
$$\mathbf{s} = \left[|V_p|, \theta_{V_p}, |I_p|, \theta_{I_p}\right]^T. \quad (5)$$
After defining the state, we define the action space of the controlling system that suits the system and can deliver the best results. Practically, the opening of the recloser is usually triggered by faults and subsequent to the series of pre-defined sequence[2]. Controlling of the recloser open is not the focus of this paper since our goal is to mitigate the TOV using proper recloser control. Therefore, we assume that the opening of reclosers is taken care of by the conventional fault detection method and the pre-defined sequence. Thus, due to the simplicity of the control task, we select a binary action space $a \in \{0,1\}$. Here, 0 indicates that no reclosing is required, whereas 1 indicates there is a reclosing action. It is necessary to remind the reader that there is an essential dimension of the action – time, which is the key to a successful reclosing.

Since the RL control agent learns through its special "feedback" – reward to improve its performance, it is important to design the reward mechanism that captures the key task sequence and maximizes its accumulative reward from the initial state to the terminal state (one episode). Therefore, we attempt to design a reward function that makes the agent learn the optimal time to reclose in the continuous state space. To achieve that, the reward function should evaluate the voltage deviation upon reclosing and consider the reclosing dead time. Consequently, for each time step t and the jth agent, we have:
$$\begin{aligned} R^j_{tovi,t} =& \alpha - \beta \cdot B_{\text{RisingEdge}} \cdot [|V_{p,t}| - V_{ref,t}]_+ \\ & - \zeta \cdot [t_{S_R=0} - t_{TH}]_+, \end{aligned} \quad (6)$$
where α, β, and ζ are adjustable scaling factors. Their values are adjustable in a specific case. The value of α determines the highest attainable reward. $B_{\text{RisingEdge}}$ is the signal bit that becomes high only when it captures the rising edge of the

[2]Electronically controlled reclosers are usually set to trip two to three times, using a combination of fast and slow time-current curves [31].

recloser j's status (changes from open (0) to close (1)). While $t_{S_R=0}$ is the time duration that recloser remains open, and t_{TH} is the allowable recloser opening time threshold that is usually the recloser dead time. The mathematical operator $[\cdot]_+$ keeps the value inside the bracket unchanged when it is non-negative, and output zero when it is negative. β and ζ denote the extent of punishment on TOV and reclosing delay. Mathematically, $R^j_{tovi,t}$ is proportional to the voltage deviation at time t from the customer-defined reference voltage, V_{ref}. Whereas, the task sequencing can be achieved by enabling the model to learn on the number of distinct action sequences.

Furthermore, it is beneficial to have a reward that evaluates the overall performance at the end of the episode. Thereby, we design the end of the episode reward R^j_{ee}:

$$R^j_{ee} = -\theta \cdot [N_{\text{Reclose}} - N_{\text{pre-defined}}]_+, \tag{7}$$

where θ is a scaling factor, N_{Reclose} and $N_{\text{pre-defined}}$ are the number of reclosing over one episode and the pre-defined number of tripping programmed in the recloser. Thus, the reward function in one episode becomes:

$$R^j = \sum_{t=1}^{T} R^j_{tovi,t} + R^j_{ee}. \tag{8}$$

Given the temporal characteristics of the reclosing task, we provide a time horizon for the task sequence in the left of the recloser controller diagram in Fig. 4 (the blue box that takes the input of the voltage and current and output of the reclosing knowledge and action). Following the time series $t_1, t_2, \cdots, t_n$, the reward comprises two parts, the instantaneous temporal sequence reward ($R^j_{tovi,t}$) and the reward at the end of the episode (R^j_{ee}). In fact, the second term in (6) pushes the model to learn the best time to reclose; whereas the third term in (6) helps the agent avoids not closing at all. To help the reader better understand the outcome of the time sequence mechanism, we assume the agent has learned "well" enough and made sure (a) the resulting voltage after reclosing is equal to V_{ref}, (b) no delayed tripping is observed, and (c) the number of reclosing matches the pre-defined value. Then, we plot Fig. 5 to show the reward with and without the time sequence design. Over the five recloser operations, the time sequence design reward captures all five reward increasing opportunities, while the one without this design can hardly do it. Since (6) indicates that optimal reward will be α which may last for Δt time, the discounted reward for each reclosure operation (reclose, wait for Δt and open again) will be bounded at $\alpha\Delta t$ when time sequences are not considered. Whereas a time sequence based reward can capture the incremental reward with increasing reclosure operations, as shown in Fig. 5c.

C. Infeasible Action Space Elimination for Fast Learning

With the time dimension considered, the action space is immense. To have a working algorithm, we need to remove most of the infeasible action space to make sure of the performance and efficiency. A generalized DQN algorithm usually solves problems or games that do not contain the time dimension. However, in this particular issue, after investigating

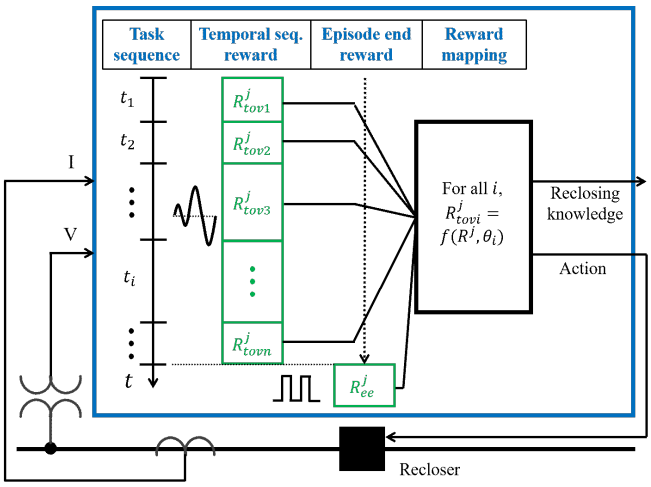

Figure 4. The schematic diagram of the proposed method's reward design.

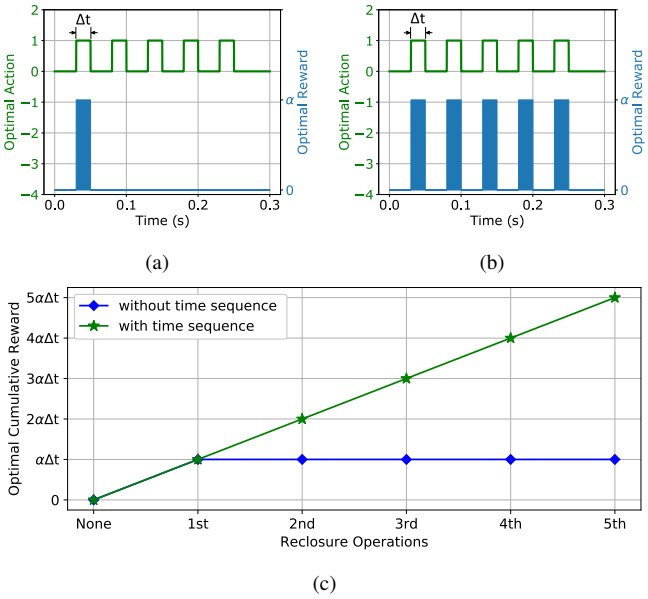

Figure 5. An example of the achieved optimal reward. (a) and (b) show the action reward pairs without and with time sequence based reward design for five breaker operations, respectively. (c) depicts the time sequence based cumulative reward goes on increasing with an increasing number of operations.

the DQN algorithm in Section III-A, we introduce the time dimension to embed the physical law into the algorithm – eliminating the physically infeasible region and enhancing the exploitation in the physically feasible region. It makes sure that we can push up the probability of action, according to the time sequence, from a state if this action is better than the value of what we should get from that state. We now redefine the probability ε in Section III-A as follows:

$$\varepsilon_t = \begin{cases} \varepsilon_0(t), & t = (\text{tr}_i, \text{tr}_i + n/f), \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

where $\varepsilon_0(t)$ denotes the base exploration rate, which is a function of time. tr_i denotes the pre-defined opening time of sequences. f is the grid frequency and n/f confines the exploration within n cycles. The agent's timer is on as long as a fault is detected.

Traditionally, the agent explores the action space from the first time step to the last one. Whereas it is not necessary most of the time if the agent wants to achieve a reduced resulting TOV. For instance the actions taken before the fault or in between two pre-defined trips are dispensable. Therefore, we conceive the notion of restricting the exploration to the time sequences where the action is required. Such a prior domain knowledge can help to gain higher rewards even in the initial few episodes. Hence, we align the temporal reward design with the temporal action likelihood. Assuming $P(a_t)$ as the prior distribution for the possible actions

$$P^*(a_t) = \begin{cases} P(a_t), & t \in \text{applicable time sequences}, \\ 0, & \text{otherwise}, \end{cases} \tag{10}$$

where $P^*(a_t)$ is the probability distribution of taking possible actions for the appropriate time sequences that exploration is needed. Such a formulation incorporates physically feasible interpretation into the model's MDP probability change. For a breaker control problem, the probabilities of having specific control actions may impact the performance mainly by restricting the exploration to a suitable temporal region and selecting appropriate probabilities of on or off actions for the breaker. So, we can perform an extensive analysis to show what probability distributions are reasonable. Therefore, we start by selecting off and on status completely randomly, i.e., both with 0.5 probability. Then we will keep on increasing the probability of occurrence of status on since the breaker is expected to remain on for more number of steps once it is reclosed. The pseudo-code is shown in Algorithm 1.

D. Post-learning Knowledge Transfer

The transferability of RL and other machine learning control methods is sometimes questioned by researchers, since, unlike deterministic control, machine learning control needs to tune its parameters based on case-specific training. This is not efficient. To overcome this issue, we adopt an approach of fitting a polynomial line $R_f \in \mathbb{R}^n$, where n is the degree of the polynomial, with reward parameters using an evaluation reward $R(S_i, A_i)$. The degree of the polynomial is a hyper-parameter which affects the speed of training:

$$\begin{aligned} R_f = \theta_0 + \theta_1 R(S_1, A_1) + \theta_2 R^2(S_2, A_2) + \cdots \\ + \theta_n R^n(S_n, A_n), \end{aligned} \tag{11}$$

where θ_i is the coefficient of the ith polynomial term. Such a polynomial function can be fitted through least square-based regression. The schematic diagram of this idea is presented in Fig. 6. We save the parameters of the reward function for the transfer learning process whenever there is a need for a new task sequence to be learn. Such a process enhances the adaptability of the model and is not restricted to only a particular environmental setting.

Algorithm 1: Deep Q-learning for Recloser Control Agent

Initialize experience buffer $\mathscr{D}$ to capacity N;
Initialize action-value function Q with random weights;
Initialize $P(a_t)$ with prior knowledge;
for *episode*=$1, E$ **do**
  Initialize sequence $s_1 = \{x_1\}$ and pre-processed sequenced $\phi_1 = \phi(s_1)$;
  for $t = 1, T$ **do**
    With probability ε_t, set $P^*(a_t) = \begin{cases} P(a_t), & t \in \text{applicable time sequences} \\ 0, & \text{Otherwise} \end{cases}$
    select a_t with probability $P^*(a_t)$; otherwise select $a_t = \max_a Q^*(\phi(s_t), a; \theta)$;
    Execute action a_t in emulator and observe reward r_t and image x_{t+1};
    Set $s_{t+1} = s_t, a_t, x_{t+1}$ and observe reward r_t and image x_{t+1};
    Store transition $(\phi_t, a_t, r_t, \phi_{t+1})$ in $\mathscr{D}$;
    Sample random minibatch (with size M) of transitions $(\phi_j, a_j, r_j, \phi_{j+1})$ from $\mathscr{D}$;
    Set $y_j = \begin{cases} r_j, & \text{for terminal } \phi_{j+1} \\ r_j + \gamma \max_{a'} Q(\phi_{j+1}, a'; \theta), & \text{for non-terminal } \phi_{j+1} \end{cases}$;
    Perform a gradient descent step on $(y_j - Q(\phi_j, a_j; \theta))^2$ based on (3)
  end
end

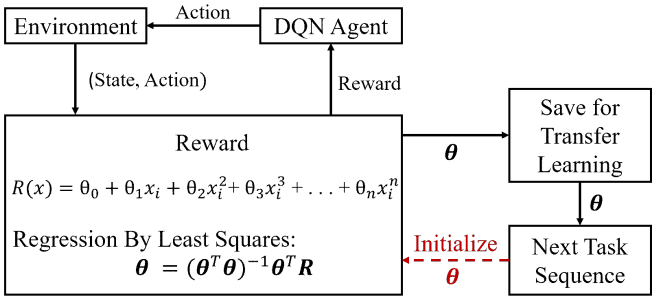

Figure 6. Scheme of learning reward function along with the agent by fitting the reward to a polynomial function.

IV. Numerical Results

A. Benchmark System

The proposed method is extensively tested in various systems. In this section, we present the results for a generalized benchmark system, as shown in Fig. 7 (refer to Appendix A). This system is a 12 kV, 100 MVA feeder. Meanwhile, the feeder circuit has 2-types of cable: (1) 750 Copper, XLPE, 15 kV 100% insulated, 26 – #22 wire shield, jacketed, and (2) 750 Aluminum, XLPE, 15 kV 100% insulated, 12 – #12 concentric neutral, jacketed. The feeder duct bank uses 3 inches PVC conduits arranged horizontally, concrete encased, burial depth 48 inches. Tests include different load conditions, source parameter change, and frequency oscillation, etc. The loads can be capacitive (C), inductive (L), resistive (R), or any of their combination. The cable is represented as a (1) a distributed line model and (2) a frequency-dependent π model using realistic underground cable parameters, the data of which is supplied and validated by our industrial partner and is shown

in Appendix A.

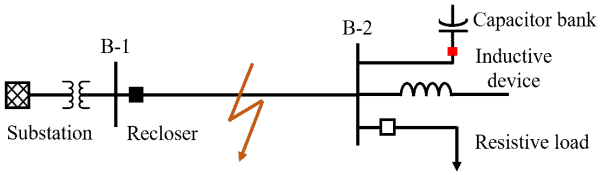

Figure 7. Benchmark system. 12 kV, 2.5 miles long underground cable, with a capacitor bank and a 8 MW load and capacitor bank at the feeder end.

With the benchmark model, we first evaluate the impact of different load types on TOVs. As shown in Table III, the load types of C and LC are two significant causes of cable TOVs. They are, in reality, the capacitor bank and the inductive loads, including transformer connected to the cable. With a decreased L or increased C, the maximum TOVs tend to increase, since the load becomes more and more capacitive in nature. Furthermore, for only load type C we observe the highest maximum TOVs, because upon reclosure the voltages are held at high values by the charged capacitor and there is no alternate route to discharge. The results also indicate that resistive load serves as the drain of the trapped charge in the cable; therefore, the TOVs are hardly observed.

Table III
IMPACT OF DIFFERENT LOAD TYPES ON TOVS.

Load Type			TOV	Underground Line
Resistive	Inductive	Capacitive		Max. Value of TOV (pu)
On	On	On	✗	-
On	Off	On	✗	-
Off	On	On	✓	1.5
Off	Off	On	✓	2.2
Off	On	Off	✗	-
On	On	Off	✗	-
On	Off	Off	✗	-

Additionally, the TOVs have large deviations when switching off the loads due to possible restrikes. Therefore, this can also be one of the reasons for causing detrimental TOVs. To study such a phenomenon of load switching due to restrikes and develop a deeper insight into the matter, we expanded the switching scenarios with rigorous experimentation to identify the highest TOV values upon multiple restrikes. Results are presented in Table IV. Our analysis shows that there is a high TOV when the load is shed without losing capacitor banks. The controller can also be designed to mitigate such TOVs.

Table IV
IMPACT OF LOAD SWITCHING ON TOVS.

Before Switching			After Switching			TOV
R	L	C	R	L	C	(pu)
On	On	On	Off	Off	On	1.54
On	On	On	Off	On	On	1.23
On	On	On	Off	On	Off	1.49
On	On	On	Off	Off	On	1.20
On	On	On	On	Off	Off	1.05

B. Overall Learning Curve by Using the Temporal Sequence Reward Mechanism and Hyper-parameter Selection

With the proposed temporal sequence reward mechanism and Deep Q-learning algorithm in Section III-B and III-C, we achieve the learning curve, as shown in Fig. 8. By looking at the average reward, it shows that the agent has many attempts to explore the optimal control action that accumulates the rewards. Some of the episode rewards are high, and some are low. A breakthrough is not realized until the episode number turns 200. After that, the agent keeps on refining its policy to improve its learning. Although the average reward gets a bit low at episode $550-750$, the agent manages to get rid of some low-performance policies and fulfill a higher reward after episode 750. Next, we will explain the way of hyper-parameter selection that helps achieve what we get.

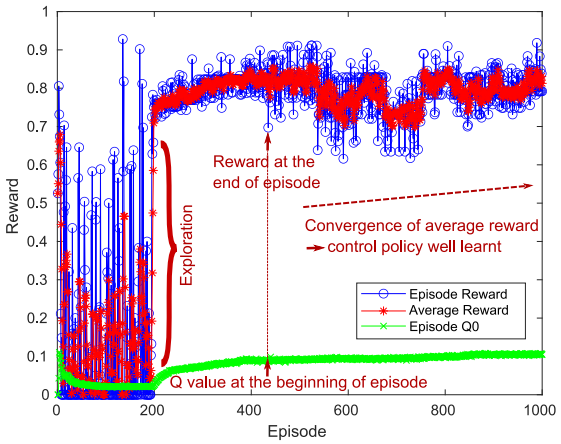

Figure 8. A learning curve that shows the individual episode reward, average reward, and Q value at the beginning of each episode named Episode $Q0$.

1) Discounting Factor (γ): Fig. 9a illustrates the effect of discounting factor on the reward. Intuitively, a value that gives the highest bounded reward, will be a fair discounting factor. But, there is a need for enough exploration as well; therefore, a discounting factor with an intensive exploration of the space while achieving a high reward would be preferable. We apply a discounting factor of 0.95 in this study so that a fair compromise is achieved between the mean value of reward and the exploration that can be shown as the standard deviation of the discounted reward values for all episodes.

2) Epsilon (ε): The exploration and exploitation are controlled by the ε value in the epsilon-greedy algorithm. By progressively increasing the epsilon from 0.85 to 0.99, we choose 0.90 as its optimal value with Fig. 9b since it shows the maximum reward achieved. It is noteworthy that increasing the epsilon further increases the likelihood of reaching a local minimum. That is why we did not adopt a higher ε that has a higher maximum reward.

3) Decay Rate: The decay rate in the epsilon-greedy algorithm is analyzed in Fig. 9c which indicates the behavior of reward by increasing decay rate value from 0.004 to 0.01. We conclude that the optimal decay rate value would be 0.005, because the mean reward is highest at that point without compromising much on the exploration. However, most exploration is shown as the standard deviation at 0.0045, but it never achieves the maximum possible reward, so its mean is very low as compared to that of the mean at the prescribed decay rate of 0.005.

4) Smoothing Factor (τ): Such a factor varies with respect to the reward value. The value of mean reward is high when a

smoothing factor of 0.01 is selected, and the standard deviation is maintained relatively high too. Both considerations are key to selecting a parameter, since we aim to maximize the reward expectation, meanwhile provide enough exploration space.

5) Experience Buffer ($\mathscr{D}$) with capacity N: Since we use experience replay to predict the value function, the size of the experience buffer needs to be decided to converge the learning model to achieve high rewards. Fig. 9e shows the relationship of reward with experience buffer. We propose to use $100,000$ as the optimal value of experience buffer, since it has a significant standard deviation to allow random exploration and achieve high reward simultaneously.

6) Minimum Batch Size (M): The minimum batch size determines the dataset to be fed to the neural networks for their learning. Fig. 9f indicates that the maximum exploration has been achieved when the size is 256 bytes. However, 512 bytes deliver a high mean reward, but the exploration is insufficient. Additionally, $1,024$ bytes result in fairly reasonable exploration with high mean but will consume too much memory, which increases the computational time and is undesirable.

The proposed learning agent is trained for different X/R ratios of the source, which impacts the cumulative reward obtained by the learning agent. The X/R ratio is varied from 8 to 20 with a step size of 4, keeping into consideration the realistic X/R ratios in a distribution network. We notice that the maximum peak TOV can reach up to 2.2 pu when the X/R ratio equals to 12. The results of mean, maximum, and standard deviation (Std.) of the reward vectors upon complete training for each sample are tabulated in Table V. The results indicate a high mean and maximum reward in all cases. Interestingly, an X/R ratio of around 12 for the system under discourse gives the highest mean and maximum reward values with the least standard deviation.

Table V
EFFECT OF CHANGE OF X/R RATIO OF THE SOURCE ON REWARD OF THE AGENT.

Source X/R Ratio	Mean Reward	Maximum Reward	Std. Reward
8	1745	3156	1164
12	2548	3311	669
16	1790	3244	1254
20	2456	3201	770

Note: the highlighted row indicates the case where a highest TOV of 2.2 pu is reached under this case study.

C. Fast Learning Curve with Infeasible Region Eliminated

We propose to eliminate the region where a particular action is infeasible from the exploration by implementing a carefully designed varying probability approach. We provide the validation of that concept in Fig. 10, which shows that we ensure a faster convergence by embedding domain knowledge in the exploration process. When the infeasible actions are not eliminated, it takes about 200 more episodes for the agent to realize a significant reward increase. Interestingly, the stable region in the middle of the learning curve without eliminated infeasible actions is even lower than the one with eliminated infeasible actions. The former takes 900 episodes to achieve the latter's reward that takes less than 200 episodes. At around

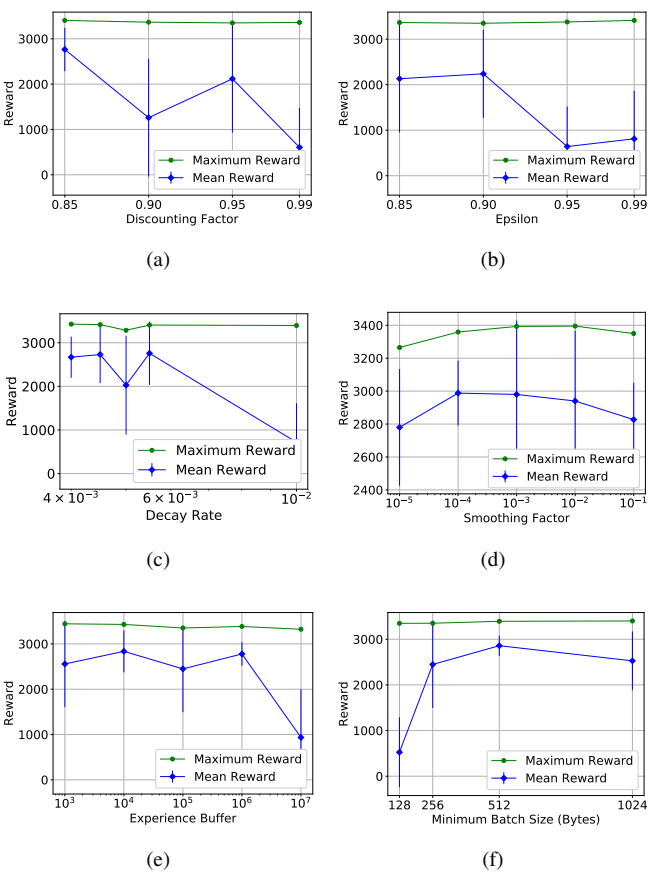

Figure 9. Effects of different hyper-parameters on the breaker controlling reward to gain insight into the problem from four different perspectives.

the 700$^{\text{th}}$ episode, the average award is boosted again with the proposed infeasible region elimination method.

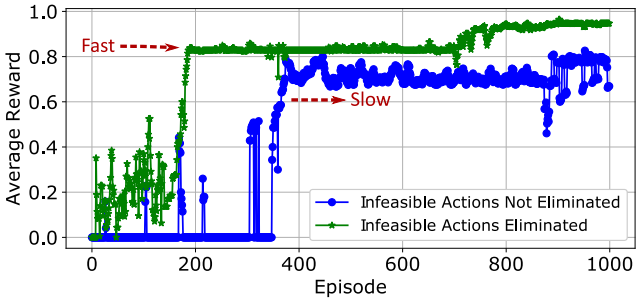

Figure 10. Learning curve comparison with and without the infeasible action eliminated. It indicates the effectiveness of eliminating the unnecessary actions by choosing a suitable probability distribution for actions while exploration.

D. Efficient Knowledge Transfer for Method Generalization

We aim to boost the learning process further to make the proposed method adaptive and general. In our application, there are multiple time sequences need to be learned by the model. Table VI illustrates that a flat start model without knowledge transfer requires many numbers of episodes to gain an average reward higher than 0.7. To ameliorate such a situation, our knowledge transfer method helps to reduce the number of episodes, since it has the capability of retaining the reward information from the past time sequences. We observe

that, with transfer learning, 261 episodes are required to gain a reward above the normalized reward of 0.7, as compared to 394 episodes with the approach of flat start, when the model is learning on the first two time sequences. For the first three time sequences, we require 289 episodes in comparison to 682 episodes. Hence, such a method of transferring reward knowledge supports the training time reduction significantly. Moreover, the generalization of reward parameters also helps in systems with other configurations to enable the reward knowledge transfer.

Table VI
EFFECT OF TRANSFERRING POST LEARNING KNOWLEDGE.

Comparison	# of episodes taken to reach average reward of 0.7		
	1 time seq.	2 time seq.	3 time seq.
Flat Start	237	394	682
Transfer Learning	212	261	289

V. PERFORMANCE COMPARISON WITH OTHER METHODS

The temporal sequence-based RL technique provides a framework to learn optimal breaker reclosure time that helps ameliorate the TOV. There have been efforts in the past to accomplish such a task. One traditional method is to reclose whenever the source side voltage crosses zero value. This zero-crossing method is easy to implement in a recloser but not effective. Therefore, we compare our proposed method with another controlled switching scheme in [25]. We call this scheme a method of half of the peak voltage, because its closing operation is performed at the instant of $+V_{max}/2$ of the source side voltage if the polarity of trapped charge is positive, and at the instant of $-V_{max}/2$ if the polarity of trapped charge is positive negative. Interested readers can refer to the cited paper to understand the mathematical formulation and the advantage of this application. Fig. 11 depicts a comparison of the proposed methodology with both of the previously adopted methods. That comparison is drawn by varying the line lengths from 1 mile to 3.5 miles with an increment of 0.5 mile, and measuring the rms voltage at the beginning of the cable. It clearly indicates that the proposed method outperforms the past techniques because the measured TOVs are the least.

Additionally, Fig. 11 illustrates a key observation about the relationship between line length and TOVs. It can be visualized that as the line length increases, the TOVs tend to decrease. Such a phenomenon is due to the progressive addition of resistance that is responsible for consuming the energy (due to the trapped charge) at reclosure operation.

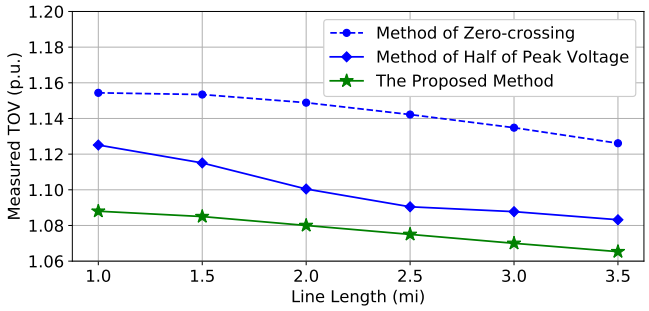

Figure 11. Comparison between proposed methodology and past methods.

VI. DISCUSSIONS

The realistic transient overvoltage should consider the modeling of restrikes/prestrikes, capacitive current, inductive current switching, the structure of the network, system parameters, whether or not virtual chopping takes place, chopping current, the instant of opening, and resonance phenomena, etc. Indeed, they are very challenging issues in transient overvoltage modeling.

Although it relates to Transmission, TOV calculations are used to determine Minimum Approach Distance (MAD) for the work rules required by the National Electrical Safety Code (NESC) and OSHA (1910.269(l)(3)(ii)). It is also our understanding TOV is dependent on the line design & operation. The work in [32] provides guidance on TOV factors and methods for control. OSHA 1926 Table 5 in Appendix A to Subpart V, provides TOV values based on various causes. Restrikes can influence TOV, but the Industry generally believes, proper periodic breaker maintenance limits the likelihood of restrikes. We assume the periodic maintenance of Distribution system breakers has a similar effect. Meanwhile, capacitor switching may have restrikes, but our paper topic concerns feeder breaking reclosing, while attempting to clear a fault. During this time the state of a switched capacitor bank remains unchanged, as well as any other devices connected to this circuit. We also assume a circuit under a fault condition, is not really lightly loaded.

Limitations exist in the transient overvoltage modeling, but this paper has demonstrated an innovative learning method that controls the reclosing under a spectral of system complexity. It relies on reinforcement learning to explore the complicated state space in a model-free way, no matter what the restrike/prestrike model is, what the structures of the network are, what the system parameters are, and whether an additional preventive device is added. Promising results are shown in the numerical section.

VII. CONCLUSIONS

Motivated by the switching-transient-related cable failures reported by our industrial partner, we develop a recloser control method for aged and degraded cables using RL. Before applying our algorithm, we find that capacitive load or capacitor banks, as well as the combination of capacitive and inductive load, are the significant causes of cable TOV phenomena. While applying our proposed algorithm, we study the impact of hyper-parameter selection on the overall learning performance and achieve a satisfactory learning curve, for which we provide our interpretation. Through our proposed time sequence reward mechanism and infeasible action elimination strategy, a fast and efficient learning curve is depicted. Comparing with the method that does not eliminate infeasible actions, the proposed method takes only 200 episodes to realize what the method without infeasible-action-elimination achieves with 900 episodes. The proposed method is also compared with one traditional method and one recent research paper. The results demonstrate the proposed RL control method has the lowest resulting TOVs. Since high-frequency oscillation associated TOVs – occurring within several nanoseconds rise time – may play an important role,

it is important to investigate the fast transient modeling and its impact in the future.

APPENDIX A
PARAMETERS OF THE BENCHMARK SYSTEM.

Table VII
SOURCE AND LINE PARAMETERS.

Source Parameters		Line Parameters	
Voltage (kV)	12	Length (mi)	2.5
Capacity (MVA)	100	Conductor	750 MCM-AL
R (Ω)	0.2326	R (Ω)	0.3163
L (H)	0.007	L (H)	0.0026
		C (pF)	112.4

Table VIII
LOAD PARAMETERS.

C (μF)	0.05
L (H)	0.08
R (Ω)	8

ACKNOWLEDGMENT

The authors would like to thank Mr. Abhishek Yatin Jathar and Mr. Ravikumar Patel for help in collecting materials.

REFERENCES

[1] J. A. Martinez, D. Goldsworthy, and R. Horton, "Switching overvoltage measurements and simulations—Part I: Field test overvoltage measurements," *IEEE Transactions on Power Delivery*, vol. 29, no. 6, pp. 2502–2509, 2014.

[2] M. M. Adibi, R. W. Alexander, and B. Avramovk, "Overvoltage control during restoration," *IEEE Transactions on Power Systems*, vol. 7, no. 4, pp. 1464–1470, 1992.

[3] K. Malmedal, "Underground vs. overhead transmission and distribution," 2009, last accessed 25 December 2019. [Online]. Available: https://www.puc.nh.gov/2008IceStorm/ST&E%20Presentations/NEI%20Underground%20Presentation%2006-09-09.pdf

[4] Q. Cui, K. El-Arroudi, and Y. Weng, "A feature selection method for high impedance fault detection," *IEEE Transactions on Power Delivery*, vol. 34, no. 3, pp. 1203–1215, 2019.

[5] Q. Cui and Y. Weng, "Enhance high impedance fault detection and location accuracy via μ-pmus," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 797–809, 2019.

[6] T. Ohno, C. L. Bak, A. Ametani, W. Wiechowski, and T. K. Sørensen, "Statistical distribution of energization overvoltages of EHV cables," *IEEE Transactions on Power Delivery*, vol. 28, no. 3, pp. 1423–1432, 2013.

[7] H. Dommel, *Case Studies for Electromagnetic Transients*. University of British Columbia, Department of Electrical Engineering, 1983. [Online]. Available: https://books.google.com/books?id=SDVOtwAACAAJ

[8] C. M. Cervantes, I. Kocar, A. Montenegro, D. Goldsworthy, T. Tobin, J. Mahseredjian, R. Ramos, J. Marti, T. Noda, A. Ametani, and C. Martin, "Simulation of switching overvoltages and validation with field tests," *IEEE Transactions on Power Delivery*, vol. 33, no. 6, pp. 2884–2893, 2018.

[9] P. Yang, S. Chen, and J. He, "Effect of different arresters on switching overvoltages in UHV transmission lines," *Tsinghua Science & Technology*, vol. 15, no. 3, pp. 325–328, 2010.

[10] E. Lindell and L. Liljestrand, "Effect of different types of overvoltage protective devices against vacuum circuit-breaker-induced transients in cable systems," *IEEE Transactions on Power Delivery*, vol. 31, no. 4, pp. 1571–1579, 2015.

[11] J. Iler and R. McDaniel, "High-speed reclosing, switching surges, and bus differential protection security - A case study," in *Annual Conference for Protective Relay Engineers*, 2017, pp. 1–13.

[12] S. Zondi, P. Bokoro, and B. Paul, "EMTP-based analysis of pre-insertion resistor and point on wave switching methodology," in *AFRICON*, 2015, pp. 1–5.

[13] P. Mestas and M. Tavares, "Comparative analysis of techniques for control of switching overvoltages during transmission lines energization," *Electric power systems research*, vol. 80, no. 1, pp. 115–120, 2010.

[14] U. Samitz, H. Siguerdidjane, F. Boudaoud, P. Bastard, J. P. Dupraz, M. Collet, J. Martin, and T. Jung, "On controlled switching of high voltage unloaded transmission lines," *E & I Elektrotechnik und Informationstechnik*, vol. 119, no. 12, pp. 415–421, 2002.

[15] N. Kolcio, J. Halladay, G. Allen, and E. Fromholtz, "Transient overvoltages and overcurrents on 12.47 kV distribution lines: Computer modeling results," *IEEE Transactions on Power Delivery*, vol. 8, no. 1, pp. 359–366, 1993.

[16] Y. Li, J. He, J. Yuan, C. Li, J. Hu, and R. Zeng, "Failure risk of UHV AC transmission line considering the statistical characteristics of switching overvoltage waveshape," *IEEE Transactions on Power Delivery*, vol. 28, no. 3, pp. 1731–1739, 2013.

[17] K. M. Dantas, W. L. Neves, and D. Fernandes, "An approach for controlled reclosing of shunt-compensated transmission lines," *IEEE Transactions on Power Delivery*, vol. 29, no. 3, pp. 1203–1211, 2013.

[18] H. Khalilnezhad, M. Popov, L. van der Sluis, J. A. Bos, and A. Ametani, "Statistical analysis of energization overvoltages in ehv hybrid ohl–cable systems," *IEEE Transactions on Power Delivery*, vol. 33, no. 6, pp. 2765–2775, 2018.

[19] D. Barros, W. L. A. Neves, and K. M. Dantas, "Controlled switching of series compensated transmission lines: Challenges and solutions," *IEEE Transactions on Power Delivery*, 2019.

[20] F. Zhang, X. Duan, M. Liao, J. Zou, and Z. Liu, "Statistical analysis of switching overvoltages in uhv transmission lines with a controlled switching," *IET Generation, Transmission & Distribution*, vol. 13, no. 21, pp. 4998–5004, 2019.

[21] J. Zhou, Y. Xin, W. Tang, G. Liu, and Q. Wu, "Impact factor identification for switching overvoltage in an offshore wind farm by analyzing multiple ignition transients," *IEEE Access*, vol. 7, pp. 64 651–64 662, 2019.

[22] S. Yinbiao, H. Bin, L. Ji-Ming, C. Weijiang, B. Liangeng, X. Zutao, and C. Guoqiang, "Influence of the switching speed of the disconnector on very fast transient overvoltage," *IEEE Transactions on Power Delivery*, vol. 28, no. 4, pp. 2080–2084, 2013.

[23] T. Suwanasri, S. Homklinkaew, and C. Suwanasri, "Effects of system configuration on switching overvoltage and insulation strength of high voltage equipment," in *International Confernce on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology*, 2010, pp. 454–458.

[24] M. Szewczyk and M. Kuniewski, "Controlled voltage breakdown in disconnector contact system for VFTO mitigation in gas-insulated switchgear (GIS)," *IEEE Transactions on Power Delivery*, vol. 32, no. 5, pp. 2360–2366, 2017.

[25] H. Seyedi and S. Tanhaeidilmaghani, "New controlled switching approach for limitation of transmission line switching overvoltages," *IET Generation, Transmission Distribution*, vol. 7, no. 3, pp. 218–225, 2013.

[26] F. Deyhim and R. Ghanizdeh, "Insulation risk assessment of controlled switching considering pre-strike voltage and line trapped charge," *IET Science, Measurement & Technology*, vol. 13, no. 2, pp. 139–148, 2018.

[27] X. Dong, Y. Yuan, Z. Gao, C. Zhou, P. Wallace, B. Alkali, B. Sheng, and H. Zhou, "Analysis of cable failure modes and cable joint failure detection via sheath circulating current," in *IEEE Electrical Insulation Conference*, 2014, pp. 294–298.

[28] "IEEE Guide for the Design and Installation of Cable Systems in Substations," *IEEE Std 525-2016 (Revision of IEEE Std 525-2007)*, pp. 1–243, 2016.

[29] D. Ha and J. Schmidhuber, "World models," *arXiv preprint arXiv:1803.10122*, 2018.

[30] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing atari with deep reinforcement learning," *arXiv preprint arXiv:1312.5602*, 2013.

[31] P. M. Anderson, *Power system protection*. Wiley, 1998.

[32] S. Surges, "Switching surges: Part iv-control and reduction on ac transmission lines," *IEEE Transactions on Power Apparatus and Systems*, no. 8, pp. 2694–2702, 1982.