Les Cahiers du GERAD ISSN: 0711–2440

Characterizing and controlling the statistics of aggregated demand-based reserve resources

A. Abiri-Jahromi, F. Bouffard, G. Joós

G–2014–38

June 2014

GERAD HEC Montréal
3000, chemin de la Côte-Sainte-Catherine
Montréal (Québec) Canada H3T 2A7

Tél. : 514 340-6053
Téléc. : 514 340-5665
info@gerad.ca
www.gerad.ca

Characterizing and controlling the statistics of aggregated demand-based reserve resources

Amir Abiri-Jahromi[a]

François Bouffard[a]

Géza Joós[b]

[a] *GERAD & Department of Electrical and Computer Engineering, McGill University, Montréal (Québec) Canada, H3A 0E9*

[b] *Department of Electrical and Computer Engineering, McGill University, Montréal (Québec) Canada, H3A 0E9*

amir.abiri-jahromi@mail.mcgill.ca
francois.bouffard@mcgill.ca
geza.joos@mcgill.ca

June 2014

Les Cahiers du GERAD

G–2014–38

Abstract: There are few systematic methodologies capable of predicting and managing the potential of large populations of appliances working as aggregated reserve resources. For demand-side based reserves to have economic and technical value, it is essential that demand-side flexibility aggregators and system operators be able to do so unequivocally. This paper introduces an analytical approach to characterize and control statistical bounds on the potential aggregated response of populations of thermostatically-controlled loads (TCLs). First, the uncertainty associated with the instantaneous power consumption of a TCL in a population is described by a set of random variables and their statistics. TCL statistics are then employed to characterize the exploitable flexibility from a large population of similar devices. From this, a control strategy and parameters are introduced to manage the aggregated response of the TCL population in response to a control signal as well as its post-response reconnection to grid. Monte Carlo simulations are employed to validate the proposed approach for the special case of a population of electric water heaters used to provide reserve capacity.

Key Words: Aggregation, ancillary services, demand response control, demand-based reserve, statistical analysis.

Résumé: Il existe présentement peu de méthodes systématiques capables de prédire et de gérer le potentiel de réglage offert par de grandes populations d'appareils électriques. Dans le cas des réserves de contingence ou d'écrêtage des pointes, il est nécessaire que les agrégateurs de flexibilités provenant de la demande soient capables d'exécuter ces fonctions avec un minimum d'erreur afin de réaliser la valeur économique et technique de cette flexibilité. Cet article introduit une approche analytique à la caractérisation et à la commande des bornes statistiques du potentiel de réglage offert par une population de charges thermostatiques (CT). Premièrement, on quantifie l'incertitude dans la consommation d'électricité instantanée d'une CT seule via un ensemble de variables aléatoires et leurs statistiques. On calcule ensuite le potentiel pour une grande population de CT similaires. De ces modèles et calculs, on pose une stratégie de commande permettant de moduler le potentiel de réglage de la population ainsi que son comportement une fois le signal de commande retiré. On valide l'approche via simulation de Monte Carlo en utilisant le cas spécifique d'une population de chauffe-eau électriques offrant un service de réglage primaire de la fréquence.

Acknowledgments: This work was supported in part by the Natural Sciences and Engineering Research Council of Canada.

1 Introduction

The vision of demand-side participation in power system operation and control is on the verge of realization as advanced metering infrastructure, embedded control systems and advanced communication technologies are becoming ubiquitous in the power industry. In chorus, the rapid integration of intermittent and non-dispatchable renewable energy resources with energy security and environmental objectives underscore the significance of demand response as an important ingredient of the smart grid paradigm [1–4].

The proliferation of renewable energy resources in power systems compels a greater need for fast acting operating reserves to offset their often unpredictable output swings [5–8]. Although several potential candidates such as pumped-hydro power plants, batteries and flywheels have already been under study to provide operating reserve, responsive demand shows great promise as it is pervasive and can offer very fast response rates. For instance, thermostatically-controlled loads (TCLs) such as air conditioners, freezers, fridges and space and water heaters can be deployed as fast responding reserve resources [9–11]. The unique characteristic of TCLs that differentiates them from other types of loads is that they can operate as an equivalent distributed energy storage asset capable of providing operating reserve without impacting much customers' comfort and productivity. Thus, there is an increasing need to develop analytical tools with the ability to characterize the flexibility and controllability that is exploitable from a certain number of TCLs. This is especially the case if system operators need to schedule such resources as part of their operating reserve mix. Likewise, demand-side flexibility aggregators, relying on a flexibility base consisting of TCL populations, can only emerge as viable businesses if such characterizations are possible at the operations planning stage.

In the 1980s and 1990s, research on TCLs was mainly focused on direct load control for peak load shaving and cold load pick-up modeling [12–16]. Most of this research considered the average demand of a population of TCLs and ignored the uncertainty associated with their instantaneous response to a control signal. More recently, research has mainly been focusing on TCL control for providing contingency reserve rather than for peak load shaving [17–22]. It has been argued that TCLs are more suitable for supplying contingency reserve than peak load shaving since the necessary response durations are shorter in the former compared to the latter case [23, 24]. In addition, the frequency of calls for deployment of contingency reserve is much lower than that for peak shaving. So far, however, there is no systematic approach available in the public domain capable of characterizing a TCL population for its reserve capacity potential at the operational planning stage. This is to contrast with important recent work in the area from [19, 20], for example, which focuses on the real-time operations of a responsive TCL population.

This is why this paper proposes an analytical approach to characterize the potential flexibility and controllability that could be exploited from a population of N TCLs as a fast response reserve resource. The original contributions of the present work, in comparison to preliminary work in [25], are the introduction of control parameters and a control strategy to manage the aggregated response of TCLs as well as their coordinated reconnection to the grid.

The approach recognizes and takes into account the fact that the aggregated response of a population of TCLs to a control signal is intrinsically uncertain because of the stochasticity of the instantaneous demand of each individual TCL in the population. Therefore, we work to characterize analytically the uncertainty associated with the instantaneous power consumption of one TCL which is then employed to compute the uncertainty associated with the exploitable flexibility from a population of N similar TCLs. Next, we address control parameters and strategies to manage 1) the level of aggregated response of a TCL population, as well as 2) its orderly reconnection to grid. Both of these are necessary conditions to provide reserve capacity products comparable to those provided by dispatchable generation.

One key attribute of our characterization methodology is that it does not rely on monitoring and/or estimating the internal temperature of individual TCLs, an approach taken by [19, 20] for real-time control. Instead we rely on TCL on-time statistics and patterns which, we argue, make the characterization much simpler. In addition, this is an attractive approach because historical TCL power consumption records can be used to determine such on-time information.

The paper is organized as follows. Section 2 presents the framework proposed to characterize the uncertain response of one TCL. Section 3 works out a characterization of the uncertain response of an entire TCL population. In Section 4, we introduce a control approach which could be used to modulate the TCL population response as required to assemble commercially-acceptable reserve capacity products. A case study based on a population of electric water heaters (EWH) is presented in Section 5 to provide validation evidence of the analytical approach. Finally, conclusions are drawn in Section 6.

2 Characterizing the Instantaneous Demand of A TCL in a Population

Figure 1 shows the generic representation of a TCL cycle, which refers to the succession on states and off states. The duty cycle of a TCL refers to the duration of the on state over the full duration of the cycle. Fig. 1 also indicates that TCL cycles may vary as a consequence of external forcing factors. For example in the case of EWHs, hot water drawage, input water temperature and ambient temperature all influence EWH cycle duration and duty cycle. Hence, the instantaneous demand of a TCL cannot be determined without metering it explicitly. Nonetheless, it is possible to establish bounds on both the cycle duration and duty cycle, as illustrated by parameters t_{min}, t_{max} , t_{min}^{cycle} and t_{max}^{cycle} in Fig. 1, while at the same time it is possible to get those without having an explicit monitoring of the TCL temperature.

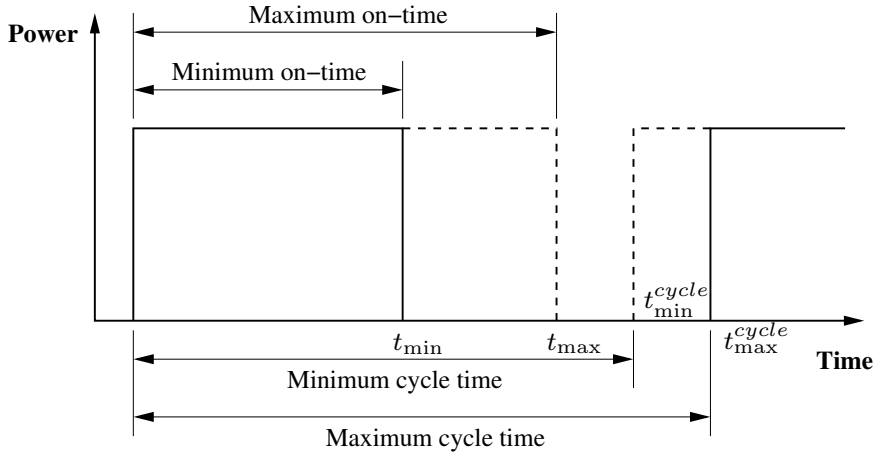

Figure 1: A generic model of a TCL cycle

Two random variables are introduced here to characterize these uncertainties. The position of a TCL within its on-off cycle is characterized by a random variable, X, in the range 0 to the normalized cycle time $t^{cycle} = 1$. We shall assume that X is uniformly distributed since an appliance could be at any point in its cycle with equal probability. Further, we denote the on-time of a TCL by random variable Y. It is noteworthy that random variable Y has only to be characterized in the time interval between $t_{min} \geq 0$ and $t_{max} \leq 1$. We note also that Y may be equal to zero if the TCL typical on-off cycle lasts longer than the period during which the TCL instantaneous consumption is assessed, typically one hour. This would be the case with electric water heaters as their on-off cycles are much longer than one hour without or with low hot water drawage [26].

The minimum and maximum duty cycles of a TCL in a population can be estimated assuming that temperature conditions, date and time of day are known. The existing physically-based models as well as laboratory analysis or field measurement data can be employed to help characterize the probability distribution of Y as well as the values of t_{min} and t_{max} as functions of time and ambient temperature. Thus, the random variable characterizing the instantaneous demand of a TCL in a population, $D(X, Y)$, is

$$D(X,Y) = \begin{cases} 1, & \text{if } X \in [0, t_{min}] \vee (X \in [t_{min}, t_{max}] \wedge X \leq Y) \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

where the demand is normalized to the appliance rated power. In the next section, we elaborate how a population of N TCLs can be leveraged by an aggregator to produce a sizeable demand response capacity as N grows large.

3 Characterizing the Exploitable Flexibility from a Population of TCLs

The context here is that of a demand response aggregator whose objective is to gather and market the equivalent capacity of a population of N TCLs. To obtain significant capacity, N has to be significant, yet that capacity remains uncertain. In this section, we use fundamental principles from probability and statistics to estimate the uncertainty as well as the expected value of the population demand response capacity.

3.1 Aggregated Demand Response Random Variable

To start, it is reasonable to assume that the available response capacity random variables for each appliance $i \in \{1, ..., N\}$ in the population, $D_i(X_i, Y_i)$, are independent and identically distributed (IID). This assumption is justified by the fact that TCLs of a similar class will share similar technical characteristics, while their on-off switching patterns will be happening independently from one another. We then define a new random variable, A_d, that represents the exploitable capacity from the population of N TCLs

$$A_d = \sum_{i=1}^{N} D_i(X_i, Y_i) \tag{2}$$

Assuming that the population of TCLs is large enough, the Central Limit Theorem (CLT) is invoked, and, consequently, we can claim that A_d is normally distributed [25]. In addition, the expected value of A_d is equal to $N\mu_d$ while its standard deviation is $\sqrt{N}\sigma_d$. Here, the parameters μ_d and σ_d respectively denote the expected value and the standard deviation of the IID TCLs, instantaneous demand random variables $D_i(X_i, Y_i)$. The parameters μ_d and σ_d are derived in Appendix A.

3.2 Uncertainty Characterization of the Aggregated Demand Response Capacity

The uncertainty associated in using the expected value of A_d, $\hat{A}_d = N\mu_d$ as an estimate of the available demand response capacity can be asserted by determining confidence intervals about the estimate. Given an actual realization of the random variable A_d, this realization should lie within the interval $\hat{A}_d \pm \nu_\gamma\sqrt{N}\sigma_d$ with probability (confidence coefficient) γ [27], *i.e.*,

$$P\{N\mu_d - \nu_\gamma\sqrt{N}\sigma_d < A_d < N\mu_d + \nu_\gamma\sqrt{N}\sigma_d\} = \gamma \tag{3}$$

where, $\nu_\gamma = \sqrt{2}\ \mathrm{erf}^{-1}\gamma$.

Thus, the statistical bounds on the exploitable flexibility from a population of N TCLs are characterized in (3) by the mean and variance of the instantaneous demand of a typical TCL in the population. It is noteworthy that the mean and variance of a TCL demand only depends on the mean of Y, the on-time duration, as derived formally in Appendix A. As a result, the exploitable flexibility of a population of N TCLs is characterized by estimating the mean of Y for a typical TCL.

4 Controlling the TCL Population Response

As mentioned already, the goal for demand response aggregators is to assemble capacity products to be used chiefly by transmission system operators as technically equivalent substitutes for generation-based reserves. To achieve this goal, the aggregator has statistical measures, as outlined above, to assess the flexible capacity potential of its TCL population and its uncertainty. It should also have ways to modulate the overall response (scale and timing) of the population in case capacity is indeed called. First, the aggregator should be able to modify the scale of the response. This is easily done through modifying the actual number of TCLs (*i.e.*, N)

being called to deliver consumption reductions.[1] Second, the aggregator should be in a position to have control over the duration of the response and over the return to "normal" operation of the TCL population.

This section deals with the first goal. Demand response resources, especially those based on TCLs, can have very short response times to a demand reduction signal. In fact, their lack of mechanical inertia makes them ideal to provide fast action reserves to offset short-term wind power variability, for example [28]. However, unlike generation-based reserves, which rely on local primary energy stockpiles to back potential electricity deliveries, TCLs called in to deploy reserves have to eventually replenish themselves with electricity so to continue on with their primary customer-driven mission. Regulating this energy payback phenomenon is critical for aggregators especially as reserve deployment duration increases [29]. The criticality comes from the need to avoid demand peaks associated to appliances reconnecting simultaneously. Moreover, system operators, in a desire to avoid post-demand response spikes, may also request a smooth and gradual return to post-response operation.

To this end a simple control strategy is proposed to regulate the aggregated response of a population of TCLs as well as its post-response reconnection to the grid. The strategy works based on the statistical knowledge of the flexibility achievable from N TCLs determined in the previous sections and which is realized through the control logic embedded in each TCL. In practice, the aggregator would have to embed controllers capable of implementing this strategy, and it should be able to act upon its parameters periodically to ensure appropriate population-wide behavior.

4.1 Control Logic to Modulate Response Magnitude

Figure 2 shows the state diagram of the TCL controller proposed here. State A is the default state in which a TCL would operate normally. Upon reception of an activation signal from the aggregator, system operator or even through the supply frequency or voltage, the controller enters State B. If the TCL is on upon entering State B and has been on for at least T_{on}^{min} then it is switched off and enters State C. Otherwise, the TCL returns to State A and will continue to cycle with State B until the activation signal is removed or the minimum on-time requirement is met. The parameter T_{on}^{min} is a control parameter introduced in the controller logic to affect the aggregated response of the population by modifying the statistics of the random variable, Y. Further, this minimum on-time requirement ensures that a TCL, once it is turned on, can store up a minimum amount of energy to guarantee customer comfort and productivity if requested to turn off.

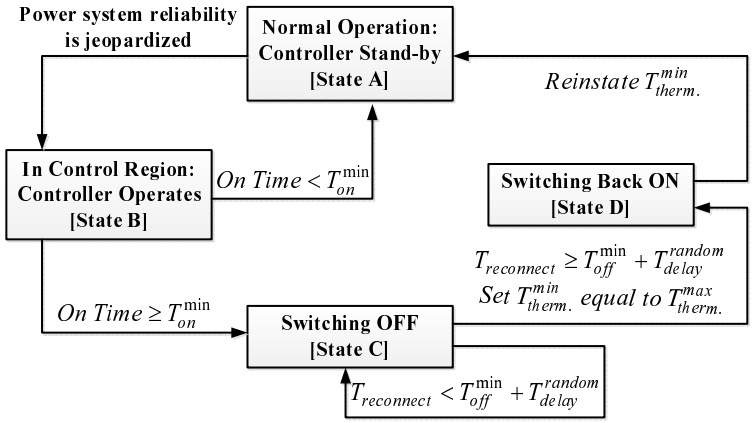

Figure 2: State diagram of TCL controller

Through the application of a minimum on-time requirement for the entire TCL population, the functional form of the random variable $D_i(X_i, Y_i)$ defined in Section 2 has to be revisited.

[1] Decreasing N in practice would not be advisable, however, because the relative response uncertainty would grow as a result. As shown later in this section, the aggregator may have the same success by controlling the statistics of the TCL on-time, while keeping its population constant.

In the case where $T_{on}^{min} \leq t_{min}$, *i.e.* the controller minimum on-time is less than the statistical minimum on-time t_{min}

$$D^1(X,Y) = \begin{cases} 1, \text{ if } X \in \left[T_{on}^{min}, t_{min}\right] \vee (X \in [t_{min}, t_{max}] \wedge X \leq Y) \\ 0, \text{ otherwise} \end{cases} \tag{4}$$

and when $t_{min} < T_{on}^{min} < t_{max}$ *i.e.*, the controller minimum on-time is greater than the statistical minimum on-time t_{min}

$$D^2(X,Y) = \begin{cases} 1, \text{ if } X \in \left[T_{on}^{min}, t_{max}\right] \wedge X \leq Y) \\ 0, \text{ otherwise} \end{cases} \tag{5}$$

In (4) and (5), D^1 and D^2 indicate whether or not a TCL would respond when needed considering the control parameter T_{on}^{min}. Given its value, it is possible to calculate μ_r and σ_r, which are, respectively, the expected value of the appropriate choice of D^1 or D^2 and its standard deviation. The statistics are calculated formally in Appendix B.

As with the original population-wide metric A_d, we can also determine the overall response potential under the control of T_{on}^{min}. We define $A_r = \sum_{i=1}^{N}[D_i^1(X_i,Y_i) + D_i^2(X_i,Y_i)]$ and its estimator $\hat{A}_r = N\mu_r$ for which we can provide confidence intervals as in (3).

The estimation of the aggregated response of a TCL population and the uncertainty associated with its response depend primarily on the on-time random variable Y of each appliance in the population (see Appendix B). The parameters t_{min} and t_{max} and the probability distribution of Y are specific to individual TCLs and their usage pattern and thus are uncontrollable. On the other hand, T_{on}^{min} can be adjusted by the aggregator to modulate the response of the population. In fact, by increasing T_{on}^{min} the aggregator is working to decrease the expected value of the random variable $D_i^1(X_i,Y_i) + D_i^2(X_i,Y_i)$ for each $i \in \{1,...,N\}$. This happens by forcing TCLs to stay on longer in the presence of an aggregator activation signal.

Moreover, T_{on}^{min} can be used to guarantee a minimum TCL primary mission performance during demand response activation periods. There is an obvious trade-off between the ability to sustain response over time and the magnitude of the population-wide response. Shorter T_{on}^{min} can free up more instantaneous capacity to be sold as reserve by the aggregator. However, the response might not be sustainable for very long before significant dissatisfaction among customers arises. In effect, depending upon the type of reserves an aggregator is attempting to offer, it would have to optimize over the value of T_{on}^{min} to determine the level of reserve to be offered given the duration over which the response would have to be sustained, all while not adversely affecting customers. Another approach to this problem could be to segment the appliance population and apply different T_{on}^{min} parameters to each segment thus allowing a wider array of response times and volumes. These are matters of ongoing investigation not addressed further in this paper, however.

4.2 Control Logic to Modulate Return to Normal TCL Operations

Another problem that has to be addressed is the orderly return to normal operation of a TCL population once the aggregator activation signal is lifted. A similar problem applies while the activation signal is still there, and appliances need to reconnect in order to replenish their energy supplies. This is an important issue especially in light of potential mass TCL reconnections and ensuing on-time synchronization which may occur. This is particularly an issue as the TCL population would effectively be delivering its reserve capacity at that time in response to some other system disturbance. On-time synchronization could lead to a magnification of the initial disturbance it was set to mitigate.

The solution to this problem is illustrated in Fig. 2. When a TCL has been turned off and enters State C, it is forced to stay off for at least $T_{reconnect}$ before it can reconnect. The control parameter $T_{reconnect}$ consists of two components: T_{off}^{min}, which is fixed by the aggregator and uniform across the population and T_{delay}^{random} which is an individual TCL randomly-generated reconnection time. With the objective of tapering off the reconnection of the population, the aggregator would impose bounds on the minimum and maximum of this random delay and its probability distribution. For instance, if T_{delay}^{random} is uniformly distributed on the interval between T_1 and T_2 minutes, one would expect the entire population to have returned to normal operation

at a constant rate in $T_2 - T_1$ minutes as seen in Fig. 3. In the same fashion as A_d, we can characterize the reconnection process statistically with confidence intervals

$$P\left\{\frac{N\mu_r - \nu_\gamma\sqrt{N}\sigma_r}{T_2 - T_1} < S_{reconnect} < \frac{N\mu_r + \nu_\gamma\sqrt{N}\sigma_r}{T_2 - T_1}\right\} = \gamma \tag{6}$$

where $S_{reconnect}$ is the rate of TCL reconnection.

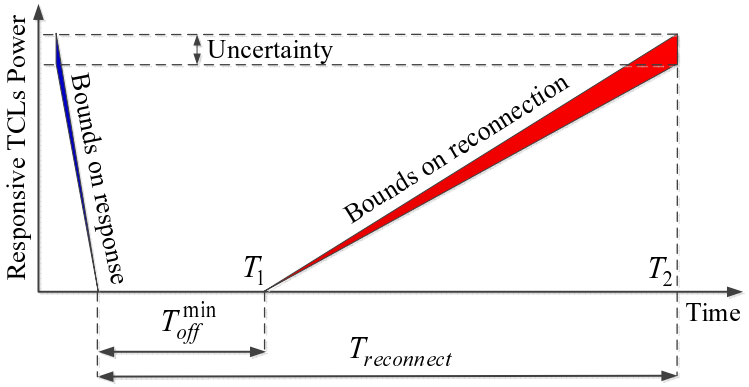

Figure 3: TCL population demand response profile

Once a TCL is allowed to reconnect (after $T_{reconnect}$), it enters State D in Fig. 2. At the same time, the TCL controller sets the TCL temperature bounds in such a way that the TCL will forcibly turn on. For instance, in the case of heating device, the minimum device temperature set point is set to its maximum set point (the opposite applies to cooling devices). This strategy ensures that the appliances primary mission is indeed satisfied while also providing better predictability over the reconnection behavior of the population as the reconnection probability becomes equal to one for each appliance that has responded. Once the maximum (minimum) appliance temperature set point is attained, the original thermostat temperature bounds are reinstated as appliances move back into the normal operation mode (State A).

Just like with T_{on}^{min}, the aggregator could set T_{off}^{min} and T_1 and T_2 (as well as the distribution of T_{delay}^{random}) by design (*i.e.*, using factory settings). This solution would not require any communications to coordinate the reconnection dynamics of the population. Nonetheless, in the case where communications are available, there would be value for the aggregator in being able to modulate over the distribution of T_{delay}^{random} and over T_{off}^{min}. That would allow the aggregator to have better control over the duration of the population response as well as over its rate of reconnection (Fig. 4). In fact, as N is set to grow, there may be a need to manage the population potential response as the load varies throughout the day and the year. A similar argument about segmenting the TCL population (as with T_{on}^{min}) can be made in order to achieve more elaborate response shapes, yet at the expense of gross volume. This aspect is outside the scope of this paper, however.

5 Case Study

The previous sections describe a general analytic approach to the characterization of the statistical bounds on the exploitable flexibility from a population of N TCLs as a reserve resource. This characterization is underpinned by the control behavior of individual TCLs in response to a demand response activation signal.

This section presents a case study of a population of N = 10,000 electric water heaters (EWH) whose consumption flexibility is used to provide reserve capacity. The objective here is to demonstrate that the analytical characterization approach, which is based on population statistics only, is appropriate to estimate the expected response of the population and its uncertainty bounds. At the same time, this should provide good evidence about how the aggregator can use individual TCL controller parameters to obtain desired aggregate reserve capacity volume, uncertainty, reconnection behavior and response duration.

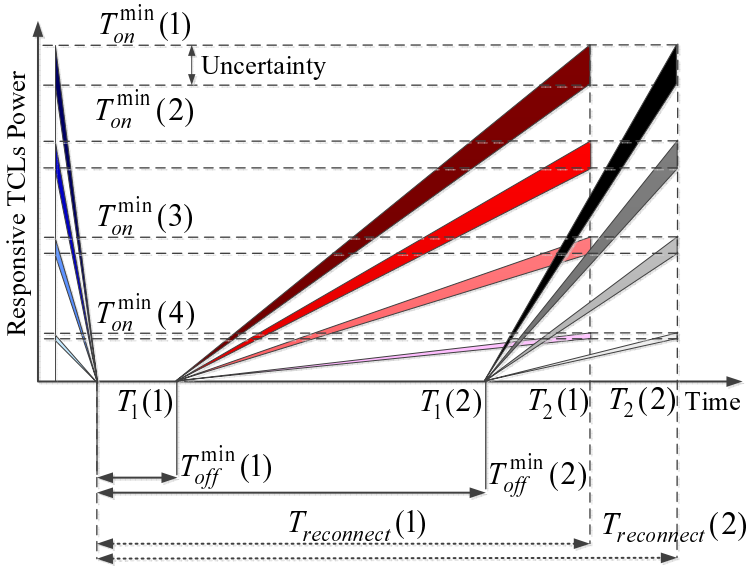

Figure 4: TCL population demand response profiles as functions of T_{on}^{min}, T_{off}^{min}, T_1 and T_2. Note that $T_{on}^{min}(1) > T_{on}^{min}(2) > T_{on}^{min}(3) > T_{on}^{min}(4)$

Appendix C derives the necessary model for the calculation of individual EWH on-time (Y) statistics based on typical EWH parameters and hot water drawage probability distribution (Tables 4 and 5, respectively, found in Appendix D). Otherwise, one could resort to laboratory data collection and analysis or to smart metering data from which EWH on-time information has been extracted from (see, for instance, [30]) on a large enough EWH population sample to obtain Y statistics for different hours and times of the year.

5.1 Characterizing the Bounds on Exploitable Flexibility from a Population of EWHs

5.1.1 Analytical Approach

The analytical approach is employed here to characterize the statistics on the flexibility that is exploitable from the 10,000 EWH population. As shown in Section 3, the flexibility exploitable from a TCL population depends on the mean value of individual TCLs Y which is calculated in (7) considering the parameters given in Tables 4 and 5

$$\mu_y = \int_0^{t_{max}} y f_Y(y)\, dy = 0.2077. \tag{7}$$

Thus, the mean μ_d and the standard deviation σ_d parameters of any $D_i(X_i, Y_i)$ are respectively 0.2077 and 0.4056 per unit (as shown in Appendix A). Further, according to the analytics derived in Section 3, the 90% confidence coefficient bounds on the flexibility achievable from this population are calculated in (8) for the hot water drawage probability distribution in Table 5 and the EWH parameters in Table 4

$$P\{2077 - \nu_{90\%}40.56 < A_d < 2077 + \nu_{90\%}40.56\} = 0.9. \tag{8}$$

From this, we find that $2010 \leq A_d \leq 2144$ with 90% certainty. Otherwise said, 90% of the time, between 2010 and 2144 EWHs out of the 10,000 in the population will respond to an aggregator activation signal.

5.1.2 Monte Carlo Simulation

In this section, we validate the analytical flexibility characterization approach by conducting a randomization experiment to obtain A_d and its statistics by simulation. Here, the temperature inside the tank and the hot water drawage are randomly drawn from the same uniform and beta distributions used in the analytical approach (with the parameters given in Tables 4 and 5).

The flexibility achievable from 10,000 EWHs is found from 10,000 Monte Carlo simulations, and the results are plotted in Fig. 5. The per-unit mean (μ_d) and standard deviation (σ_d) of the Monte Carlo simulations are 0.2067 and 0.4049 respectively. These values are consistent with the analytical results obtained previously (μ_d here is within 0.48% of the value found in Section 5.1.1 and σ_d is within 0.17% of the analytically-found value).

The 90% confidence interval and the three standard deviation boundaries obtained from the analytical approach are shown by solid and dashed lines respectively in Fig. 5 to show their close agreement with the simulation experiment. It is noteworthy that according to the well-known rule of thumb for normal distributions [27], roughly 99.7% of the samples should lie within three standard deviations from the mean since the aggregated response should have a normal distribution by the Central Limit Theorem.

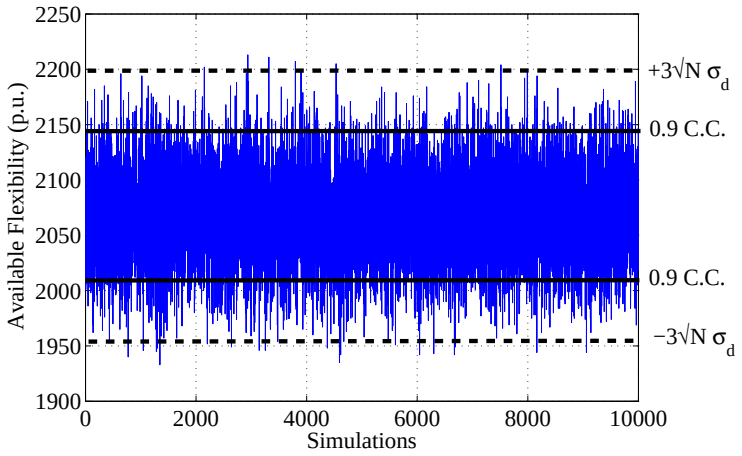

Figure 5: Representation of 10,000 Monte Carlo simulations characterizing the achievable flexibility from 10,000 EWHs

5.2 Controlling the Statistical Bounds of the EWHs Response as well as their Reconnection

5.2.1 Analytical Approach

As discussed in Section 4, the response level of the population of TCLs is controllable by adjusting the parameter T_{on}^{min}. Further, the number of EWHs reconnecting to grid at each instant of time is controllable by adjusting the parameter $T_{reconnect}$ while considering the response bounds. The aggregated response level of EWHs for three different values of parameter T_{on}^{min}, *i.e.*, 10 minutes, 20 minutes and 25 minutes, are calculated using the analytical approach and summarized in Table 1.

In addition, the rate of EWH reconnection to the grid at each instant of time, $S_{reconnect}$ as defined in (6), is calculated here for two different values of the parameter $T_{reconnect}$, *i.e.*, 25 minutes and 30 minutes. The parameter $T_{reconnect}$ consists of the fixed parameter T_{off}^{min}, which here equals 20 minutes, and the random time delay parameter T_{delay}^{random} distributed uniformly between $T_1 = 0$ and to either $T_2 = 5$ or 10 minutes respectively. We recall that T_{off}^{min} ensures that the duration of the response is minimally sustained, while the difference $T_2 - T_1$ determines the number of EWHs reconnecting to the grid at each instant of time. The results, from the analytical calculations, are summarized in Table 2.

5.2.2 Monte Carlo Simulations

Figure 6 illustrates the Monte Carlo simulation results for the previously used three values of T_{on}^{min}, (10, 20 and 25 minutes). As expected, it can be seen from Fig. 6 that the aggregated response is reduced by increasing the parameter T_{on}^{min}. It is also noteworthy that the uncertainty associated with the response reduced accordingly by increasing that parameter. This is because the number of potentially responding

Table 1: 90% Statistical bounds on aggregated response of 10,000 EWHs for different values of T_{on}^{min}

T_{on}^{min} (min)	$N\mu_r$ (p.u.)	$\sqrt{N}\sigma_r$ (p.u.)	90% A_r (p.u.)
10	1542	36.1	(1483, 1602)
20	1008	30.1	(959, 1058)
25	740	26.2	(698, 784)

Table 2: 90% Statistical bounds on the number of EWHs reconnecting to the grid

$T_2 - T_1$ (min)	T_{on}^{min} (min)	$N\mu_r$ (p.u.)	$\sqrt{N}\sigma_r$ (p.u.)	90% $S_{reconnect}$ (p.u./min)
5	10	1542	36.1	(297, 321)
	20	1008	30.1	(192, 212)
	25	740	26.2	(140, 157)
10	10	1542	36.1	(149, 161)
	20	1008	30.1	(96, 106)
	25	740	26.2	(70, 79)

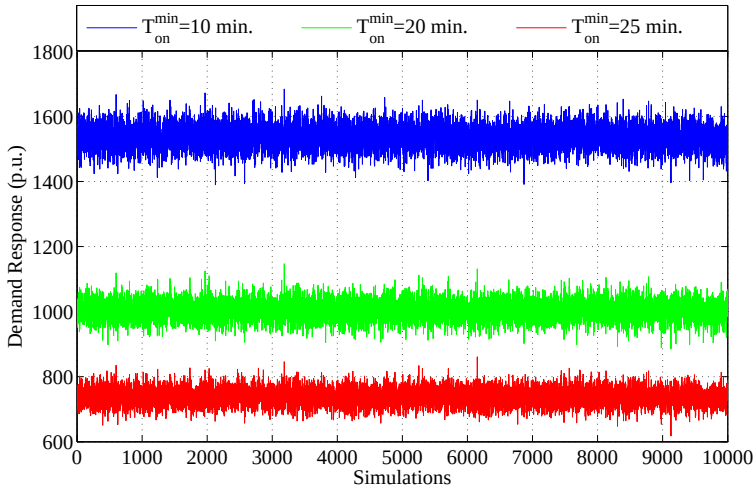

Figure 6: Representation of 10,000 Monte Carlo simulations indicating the controllability of the response level of 10,000 EWHs using parameter T_{on}^{min}

EWHs is reduced as they are forced to remain on if an activation signal is received. On visual inspection, the Monte Carlo simulation results obtained here are consistent with the numbers obtained by the analytical approach.

In addition, the EWHs reconnection to the grid is replicated using Monte Carlo simulation for T_{delay}^{random} uniformly distributed between $T_1 = 0$ and $T_2 = 5$ minutes as shown in Fig. 7. As it can be seen, the rate at which EWH reconnect to grid is controllable by the range of $T_2 - T_1$.

5.3 Employing EWHs for Primary Frequency Control

In this section, we apply the analytical characterization technique and the control of a TCL population for the provision of reserve capacity for primary frequency control. In this case, the aggregator activation signal is the grid frequency error measured by each TCL combined with aggregator-assigned cutoff frequency error thresholds. In other words, each TCL measures the grid frequency and compares it to its reference and responds if the error is over aggregator-assigned threshold, in a way similar to what is described in [18, 21].

In this case study, we assume that the cutoff frequency error thresholds of 10,000 EWHs are uniformly distributed between frequencies of 58.5 Hz to 59.5 Hz with equal spacing of 0.01 Hz (thus allocating 100

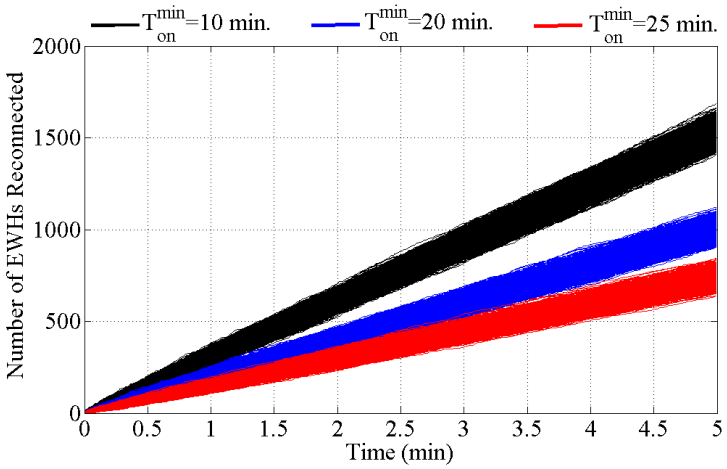

Figure 7: Representation of 10,000 Monte Carlo simulations for the EWHs reconnection to grid when the parameter T_{delay}^{random} is 5 minutes

EWHs per 0.01 Hz frequency band). Moreover, to assess the impact of T_{on}^{min} on the population response, T_{on}^{min} is varied in five minute increments from 0 to 25 minutes to show that the aggregated response of EWHs emulates a droop characteristic similar to that of conventional generating units with governor control.

5.3.1 Analytical Approach

The statistics μ_r, σ_r, the 90% confidence coefficient statistical bounds on the EWH population response over the frequency range of 58.5 Hz to 59.5 Hz, the mean droop and droop three sigma bounds are summarized in Table 3. Note that the bounds on A_r provided in Table 3 represent the number of EWHs responding to grid frequency errors at each of the 0.01 Hz frequency steps between 58.5 to 59.5 Hz as EWH population power is normalized.

Table 3: 90% Statistical bounds on aggregated response of 10,000 EWHs participating in primary frequency control

T_{on}^{min} (min)	$N\mu_r$ (p.u.)	$\sqrt{N}\sigma_r$ (p.u.)	90% A_r (p.u./0.01 Hz)	Mean droop (p.u./Hz)	$3\sigma_r$ droop (p.u./Hz)
0	2077	40.6	(14, 28)	2077	± 122
5	1809	38.5	(12, 25)	1809	± 116
10	1542	36.1	(10, 22)	1542	± 109
15	1275	33.4	(7, 18)	1275	± 101
20	1008	30.1	(5, 15)	1008	± 91
25	740	26.2	(3, 12)	740	± 79

It can be seen from Table 3 that the number of EWHs responding to frequency errors is reduced by increasing parameter T_{on}^{min} from 0 to 25 minutes. It is noteworthy that the uncertainty associated with the cumulative response of EWHs increases from frequency 59.5 Hz to 58.5 Hz by the factor of 10. This is because, the uncertainty is proportional to $\sqrt{N}\sigma_r$ and the number of EWHs participating in frequency control increases from 100 to 10,000 over the frequency range 59.5 Hz to 58.5 Hz. It can also be seen from Table 3 that the droop and droop 3σ bounds, which respectively indicate the average number of EWHs responding to frequency errors when the frequency drops to 58.5 Hz and the corresponding 3σ bounds, reduces by increasing T_{on}^{min}.

5.3.2 Monte Carlo Simulations

As it can be seen in Fig. 8, the EWH population emulates a droop characteristics similar to the conventional generating units whose slope is controllable by parameter T_{on}^{min}. The first difference between the droop

characteristics of a conventional generating unit and a TCL-based resource is that the droop in the former one is defined by a fixed value while the droop in the latter is a range as summarized in Table 3. Second, the duration that the response can be sustained by a TCL-based resource is defined by the parameter T_{off}^{min} which depends on customer comfort level that has to be satisfied. The simulation results are confirming the analytical estimations found in Table 3.

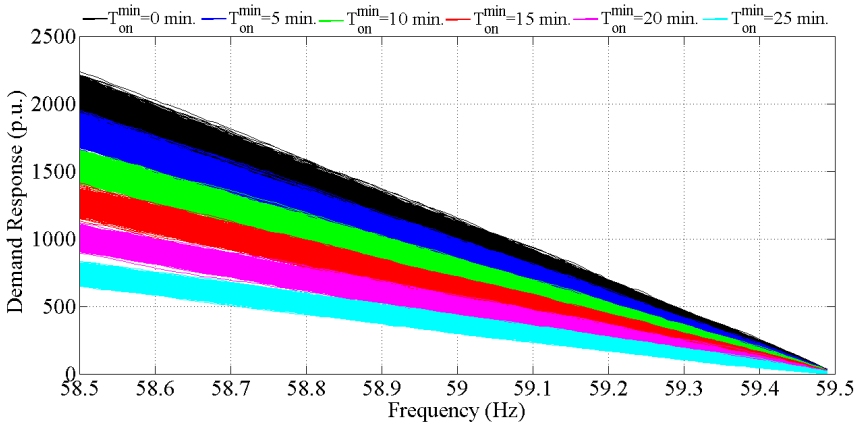

Figure 8: Representation of 10,000 Monte Carlo simulations for the cumulative response of 10,000 EWHs which are uniformly distributed over the frequency range of 59.5 Hz to 58.5 Hz with different values of T_{on}^{min}

6 Conclusion

This paper presented an analytical approach to characterize the statistical bounds on the exploitable flexibility from a population of TCLs as a reserve resource. It is revealed that the statistical bounds on the exploitable flexibility and the associating uncertainty can be characterized by evaluating the minimum and maximum duty cycles of TCLs in a population as well as the duty cycle variation probability distribution. Further, it is shown that the aggregated response of TCLs and TCLs reconnection to grid can be controlled through a control logic embedded in each endpoint TCL controller with low requirements for communications. The validity of the proposed analytical models is also investigated and verified for the special case of water heaters through mathematical modeling and Monte Carlo simulations. The results here are essential for the potential uptake of TCL-based resources in power systems operation and control. Further research should be carried out in this area to develop a methodology for scheduling and managing the response of TCL-based reserve resources.

A TCL Individual Statistics

Parameters μ_d and σ_d in (3) are calculated from

$$\mu_d = 0 \cdot P(D=0) + 1 \cdot P(D=1) = P(D=1) \tag{9}$$

where, from the definition of $D(X,Y)$ in (1),

$$P(D=1) = P(0 \le X \le t_{min}) + P\left((t_{min} \le X \le t_{max}) \wedge (X \le Y)\right). \tag{10}$$

Recalling that X is uniformly distributed over the range $[0, t^{cycle}]$, we have $f_X(x) = 1/t^{cycle} = 1$ (since $t^{cycle} = 1$), and we thus find

$$P(0 \leq X \leq t_{min}) = \int_0^{t_{min}} f_X(x)\,dx = t_{min} \tag{11}$$

$$P((t_{min} \leq X \leq t_{max}) \wedge (X \leq Y)) = \int_{t_{min}}^{t_{max}} (y - t_{min}) f_Y(y)\,dy = \mu_y - t_{min} \tag{12}$$

and thus

$$\mu_d = \mu_y. \tag{13}$$

The variance of $D(X, Y)$ is calculated as

$$\begin{aligned} \sigma_d^2 &= (0 - \mu_d)^2 P(D = 0) + (1 - \mu_d)^2 P(D = 1) \\ &= \mu_d^2(1 - \mu_d) + (1 - \mu_d)^2 \mu_d \\ &= \mu_d(1 - \mu_d) \end{aligned} \tag{14}$$

where we used the result of (9) and the fact that $P(D = 0) = 1 - P(D = 1)$.

B Statistics of Individual TCLs Influenced by T_{on}^{min}

Parameters μ_r and σ_r^2 are found analytically

$$\mu_r = \begin{cases} \mu_y - T_{on}^{min}, & \text{if} T_{on}^{min} \leq t_{min} \\ \mathbb{E}[Y | y \geq T_{on}^{min}], & \text{otherwise} \end{cases} \tag{15}$$

where

$$\mathbb{E}[Y | y \geq T_{on}^{min}] = \int_{T_{on}^{min}}^{t_{max}} (y - T_{on}^{min}) f_Y(y)\,dy \tag{16}$$

is the conditional expectation of Y given that the TCL on-time is greater than or equal to T_{min}^{on}. Here, we can reasonably assume that $T_{on}^{min} < t_{max}$. In addition,

$$\sigma_r^2 = \begin{cases} (\mu_y - T_{on}^{min})(1 - \mu_y + T_{on}^{min}), & \text{if } T_{on}^{min} \leq t_{min} \\ \mathbb{E}[Y^2 | y \geq T_{on}^{min}] - \mathbb{E}[Y | y \geq T_{on}^{min}]^2, & \text{otherwise.} \end{cases} \tag{17}$$

C EWH On-Time Modeling

A typical EWH consists of a storage tank, a thermostat, a heating element, an inlet (cold) water pipe and a hot water outlet pipe [26].

To determine t_{min} and t_{max} and the probability distribution of Y, we model the energy flows in a typical EWH. The drivers for energy flows are: 1) hot water usage followed by its replacement by cold water and 2) wall conduction losses. The first driver dominates energy flows because the time constants for energy flows via wall conduction are longer. When the EWH does have to turn on, its minimum on-time (t_{min}) is associated with the tank water temperature having dropped below the minimum thermostat set-point T_{th}^{min} without water drawage. This event activates the heating element until the water temperature rises to the maximum thermostat set-point T_{th}^{max}. Therefore,

$$t_{min} = \frac{c\rho V(T_{th}^{max} - T_{th}^{min})}{P_r} \tag{18}$$

where c and ρ denote the specific heat and density of water, respectively, V is the tank volume and P_r is the rated power of the heating element. The maximum on-time (t_{max}), on the other hand, occurs when the tank

water temperature drops below the minimum thermostat set-point concurrently with maximum hot water drawage

$$t_{max} = t_{min} + \frac{c\rho L_d^{max}(T_{th}^{min} - T_{in})}{P_r} \tag{19}$$

where, L_d^{max} is the maximum hot water drawage, and T_{in} is the inlet water temperature. This result is interpreted as the time to heat up to T_{th}^{min} a volume of L_d^{max} liters from an initial temperature of T_{in} followed by further heating time of t_{min} to bring the tank water volume up to T_{th}^{max}.

The EWH on-time random variable Y is a function of two other random variables, namely T_T and L_d. Here T_T is the water temperature measured by the thermostat when drawage starts. L_d is the total volume of water drawn which accounts for water drawn continuously before the EWH turns on and for water drawn while the EWH is on. From this, Y takes the form

$$Y = g(T_T, L_d) = \frac{c\rho V(T_{th}^{max} - T_T)}{P_r} + \frac{c\rho L_d(T_T - T_{in})}{P_r} \tag{20}$$

In a way similar to (19), Y accounts for the need to bring up the entire volume of water from T_T to T_{th}^{max} (first term) and the drawn amount from T_{in} up to T_T (second term).

In the absence of water drawage, the water temperature behaves as a uniform random variable ranging between the EWH thermostat set-points

$$f_{T_T}(t_T) = \frac{1}{T_{th}^{max} - T_{th}^{min}} \tag{21}$$

As for L_d, unlike the approaches of [19] and [26], we recall that this amount is a total volume of water drawn and not a rate of drawage (in liters per minute, for example). Therefore, the typical Markov models for hot water consumption are not appropriate here. Instead, for the sake of this study, we assume that L_d is beta-distributed for the following two main reasons. First, the beta distribution is bounded, which is the case for hot water consumption, i.e., L_d^{min}, L_d^{max}. Second, because it has shape parameters p and q, the distribution can be adjusted to capture randomness that displays both skew and kurtosis [31]. Therefore,

$$f_{L_d}(\ell_d) = \frac{\Gamma(p+q)}{\Gamma(p)+\Gamma(q)} \frac{(\ell_d - L_d^{min})^{(p-1)}(L_d^{max} - \ell_d)^{(q-1)}}{(L_d^{max} - L_d^{min})^{(p+q-1)}} \tag{22}$$

where $\Gamma(\cdot)$ is the gamma function.

We assume that T_T and L_d are not correlated at the beginning of a water drawage interval. Further, it is important to recognize that there is a minimum volume of continuous water drawage over which the EWH turns on because it brings the overall EWH water temperature below T_{th}^{min}. This volume of water, L_d^*, is such that $(V - L_d^*)(T_T - T_{th}^{min}) = L_d^*(T_T - T_{in})$, which implies that

$$Y = \begin{cases} 0, & \text{if } L_d < L_d^* \\ g(T_T, L_d), & \text{otherwise} \end{cases} \tag{23}$$

where $L_d^* = V(T_T - T_{th}^{min})/(T_T - T_{in}) \geq 0$ depends on the realization of T_T at the beginning of the drawage period. We note that in the case where $L_d = 0$ and the water temperature drops to T_{th}^{min}, (23) sets the EWH on-time to t_{min}.

We can thus infer the conditional cumulative distribution function of Y from those of T_T and L_d

$$F_Y(y|t_T, \ell_d) = F_{L_d}\left(\frac{V(t_T - T_{th}^{min})}{t_T - T_{in}}\right) \cdot \left[1 - F_{T_T}\left(\frac{VT_{th}^{min} - \ell_d T_{in}}{V - \ell_d}\right)\right], \text{ if } 0 \leq y < t_{min} \tag{24}$$

$$F_Y(y|t_T, \ell_d) = \left[1 - F_{L_d}\left(\frac{V(t_T - T_{th}^{min})}{t_T - T_{in}}\right)\right] \cdot F_{T_T}\left(\frac{VT_{th}^{min} - \ell_d T_{in}}{V - \ell_d}\right), \text{ if } t_{min} \leq y \leq t_{max}. \tag{25}$$

The probability distribution function of Y can be obtained from these to compute its mean and variance (Appendix A).

D EWH Data

The parameters of a typical EWH and the parameters of the hot water drawage beta distribution are given in Tables 4 and 5 respectively.

Table 4: Parameters of typical EWH [26]

Parameter	Value
Volume (V)	50 gal (0.1893 m^3)
Max. thermostat setpoint (T_{th}^{max})	135°F (57.2°C)
Min. thermostat setpoint (T_{th}^{min})	115°F (46.1°C)
Inlet water temperature (T_{in})	60°F (15.5°C)
Rated power (P_r)	4.5 kW

Table 5: Hot water drawage beta distribution parameters

Parameter	Value
Max. hot water drawage (L_d^{max})	25 gal (0.0947 m^3)
Min. hot water drawage (L_d^{min})	0 gal (0 m^3)
Shape parameter (p)	2
Shape parameter (q)	8
Hot water drawage mean (μ_{L_d})	5 gal (0.0189 m^3)

References

[1] U.S. Federal Energy Regulatory Commission, National Action Plan on Demand Response, FERC, Washington, DC, Jun. 2010, [On-line], Available: www.ferc.gov/legal/staff-reports/06-17-10-demandresponse.pdf.

[2] E. Hirst and B. Kirby, Load as a resource in providing ancillary services, Oak Ridge National Laboratory, Oak Ridge, TN. Tech. Rep. Jan. 1999. [Online]. Available: http://www.ornl.gov/sci/btc/apps/Restructuring/Load-as-a-Resource-APC-99.pdf.

[3] M. Milligan and B. Kirby, Utilizing Load Response for Wind and Solar Integration and Power System Reliability, NREL, Boulder, CO. Tech. Rep. NREL/CP-550-48247, May, 2010.

[4] F. Rahimi and A. Ipakchi, Demand response as a market resource under the smart grid paradigm, IEEE Trans. Smart Grid, 1(1), 82–88, Jun. 2010.

[5] Integration of Variable Generation Task Force, Accommodating High Levels of Variable Generation, Tech. Rep. NERC, Princeton, NJ, Apr. 2009.

[6] J. Undrill, Power and Frequency Control as it Relates to Wind-Powered Generation, Lawrence Berkeley National Laboratory, Berkeley, CA, Tech. Rep. LBNL-4143E, Dec. 2010.

[7] E. Ela, M. Milligan, B. Kirby, A. Tuohy, and D. Brooks, Alternative Approaches for Incentivizing the Frequency Responsive Reserve Ancillary Service, Tech. Rep. NREL/TP-5500-54393, March 2012.

[8] J. Adams et al., Flexibility Requirements and Potential Metrics for Variable Generation: Implications for System Planning Studies, NERC, Princeton, NJ, 2010. [Online]. Available: http://www.nerc.com/files/IVGTF-Task-1-4-Final.pdf.

[9] B. Kirby and M. Ally, Spinning Reserve from Supervisory Thermostat Control, ser. Transmission Reliability Research Review, US DOE, Washington, DC, Dec. 2002.

[10] D.J. Hammerstrom et al., Pacific Northwest Grid Wise Testbed Demonstration Projects: Part I. Olympic Peninsula Project, PNNL, Richland, WA, Tech. Rep. PNNL-17167, Oct. 2007. Available at http://gridwise.pnl.gov.

[11] D. Hirst, Responsive Load System, U.K. Patent GB2361118.

[12] C.Y. Chong and A.S. Debs, Statistical synthesis of power system functional load models, in Proc. 18th IEEE Conf. on Decision and Control, 1979.

[13] S. Ihara and F.C. Schweppe, Physically based modeling of cold load pickup, IEEE Trans. Power App. Syst., PAS-100(9), 4142–4150, Sep. 1981.

[14] R. Malhamé and C.Y. Chong, Electric-load model synthesis by diffusion approximation of a high-order hybrid-state stochastic-system, IEEE Trans. Automatic Control, 30, 854–860, 1985.

[15] J.C. Laurent and R.P. Malhamé, Physically based computer model of aggregate electric water heating loads, IEEE Trans. Power Syst., 9(3), 1209–1217, Aug. 1994.

[16] J.C. Laurent, G. Desaulniers, R.P. Malhamé, and F. Soumis, A column generation method for optimal load management via control of electric water heaters, IEEE Trans. Power Syst., 10(3), 1389–1400, Aug. 1995.

[17] S. Katipamula and N. Lu, Evaluation of residential HVAC control strategies for demand response programs, ASHRAE Trans., 112(1), 535–546, Jan. 2006.

[18] J.A. Short, D.G. Infield, and L.L. Ferris, Stabilization of grid frequency through dynamic demand control, IEEE Trans. Power Syst., 23(3), 1284–1293, Aug. 2007.

[19] D.S. Callaway, Tapping the energy storage potential in electric loads to deliver load following and regulation, with application to wind energy, Energy Conv. Manag., 50(5), 1389–1400, May 2009.

[20] J.L. Mathieu, S. Koch, and D.S. Callaway, State estimation and control of electric loads to manage real-time energy imbalance, IEEE Trans. Power Syst., 28(1), 430–440, Feb. 2013.

[21] A. Molina-García, F. Bouffard, and D.S. Kirschen, Decentralized demand-side contribution to primary frequency control, IEEE Trans. Power Syst., 26(1), 411–419, Feb. 2011.

[22] Z. Xu, J. Ostergaard, and M. Togeby, Demand as frequency controlled reserve, IEEE Trans. Power Syst., 26(3), 1062–1071, Aug. 2011.

[23] B. Kirby, Demand Response for Power System Reliability: FAQ, ORNL, Oak Ridge, TN. Tech. Rep. ORNL/TM-2006/565, Dec. 2006.

[24] D. Todd, M. Caufield, B. Helms, M. Starke, B. Kirby, and J. Kueck, Providing Reliability Services through Demand Response: A Preliminary Evaluation of the Demand Response Capabilities of Alcoa Inc., ORNL, Oak Ridge, TN. Tech. Rep. Jan. 2009.

[25] A. Abiri-Jahromi and F. Bouffard, Characterizing statistical bounds on aggregated demand-based primary frequency control, in Proc. 2013 IEEE PES General Meeting, Vancouver, BC, 2013.

[26] J. Kondoh, N. Lu, and D.J. Hammerstrom, An evaluation of the water heater load potential for providing regulation service, IEEE Trans. Power Syst., 26(3), 1309–1316, Aug. 2011.

[27] A. Papoulis, Probability, Random Variables, and Stochastic Processes, 3rd ed., New York, NY: McGraw-Hill, 1991.

[28] J. Apt, The spectrum of power from wind turbines, J. Power Sources, 169(2), 369–374, Jun. 2007.

[29] N. Ruiz, I. Cobelo, and J. Oyarzabal, A direct load control model for virtual power plant management, IEEE Trans. Power Syst., 24(2), 959–966, May 2009.

[30] M. Dong, P.C.M. Meira, W. Xu, and C.Y. Chung, Non-Intrusive Signature Extraction for Major Residential Loads, IEEE Trans. Smart Grid, 4(3), 1421–1430, Sep. 2013.

[31] C. Forbes, M. Evans, N. Hastings, and B. Peacock, Statistical Distributions, New York, NY: Wiley, 2010.