LBNL-5129E

ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

Variability in Automated Responses of Commercial Buildings and Industrial Facilities to Dynamic Electricity Prices

Johanna L. Mathieu, Duncan S. Callaway, Sila Kiliccote

Environmental Energy Technologies Division

August 2011

Submitted to
Energy and Buildings, in press

DISCLAIMER

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

Variability in Automated Responses of Commercial Buildings and Industrial Facilities to Dynamic Electricity Prices

Johanna L. Mathieu, Duncan S. Callaway, and Sila Kiliccote

Abstract

Changes in the electricity consumption of commercial buildings and industrial facilities (C&I facilities) during Demand Response (DR) events are usually estimated using counterfactual baseline models. Model error makes it difficult to precisely quantify these changes in consumption and understand if C&I facilities exhibit event-to-event variability in their response to DR signals. This paper seeks to understand baseline model error and DR variability in C&I facilities facing dynamic electricity prices. Using a regression-based baseline model, we present a method to compute the error associated with estimates of several DR parameters. We also develop a metric to determine how much observed DR variability results from baseline model error rather than real variability in response. We analyze 38 C&I facilities participating in an automated DR program and find that DR parameter errors are large. Though some facilities exhibit real DR variability, most observed variability results from baseline model error. Therefore, facilities with variable DR parameters may actually respond consistently from event to event. Consequently, in DR programs in which repeatability is valued, individual buildings may be performing better than previously thought. In some cases, however, aggregations of C&I facilities exhibit real DR variability, which could create challenges for power system operation.

Keywords: demand response, baseline models, load prediction, error analysis, variability, measurement & verification

I. Introduction

Buildings are becoming increasingly important as active resources that support power system operations. Though buildings have played a small role in power systems operations in the past – either with relays that interrupt power to air conditioners and water heaters [1], [2], or by "voice dispatch" of large commercial and industrial loads [3] – recent Smart Grid investments are demonstrating the potential for buildings to become grid-interactive resources that are just as controllable as – or even more controllable than – electricity generators [4].

In "demand response" (DR) programs, power system operators can achieve system-wide demand reductions by providing financial incentives for buildings to change their electricity consumption patterns through both "shifts" in energy use and load reductions, or "sheds." Buildings generally participate in DR by enrolling in dynamic electricity pricing programs or demand/capacity bidding programs. In dynamic pricing programs, buildings face high electricity prices during hours when the grid is stressed, encouraging them to shed load or shift energy use to less expensive hours. In capacity/demand bidding programs, buildings bid load reductions and, if called upon, shed load at certain times in exchange for payment. In this paper, we focus on commercial buildings and industrial facilities (C&I facilities) participating in a dynamic electricity pricing program. These facilities use the Open Automated Demand Response (OpenADR) Communication Specification [5] to receive DR event notifications from the utility, and during events they automatically execute pre-programmed DR strategies [6].

The central challenge we address in this paper is that DR parameters, such as Average Demand Shed, Rebound, Daily Peak Demand, and Daily Energy (which we define in Table 2), must be measured relative to an estimate of how much electricity a facility would have consumed in the absence of the DR event. DR parameters are computed by subtracting a counterfactual baseline from the actual power consumption of the facility. Therefore, DR parameters will exhibit variability due to both baseline model error and real variability in the facility's response. We will use the following terms throughout the paper:

1) **Unmodeled load variability**, or baseline model error, is load variability that is not captured by a baseline model and not due to a DR signal. Unmodeled load variability complicates DR programs that use baselines for financial settlement (e.g., demand/capacity bidding programs and programs in which loads participate in wholesale ancillary services markets). Moreover, even DR programs that do not use baselines for settlement (e.g., dynamic electricity pricing programs) use baselines for Measurement and Verification (M&V) and to calculate the cost-effectiveness of the DR programs [7].
2) **Real DR variability** is event-to-event variability in a facility's actual response, for example, due to building managers and/or occupants overriding pre-programmed DR strategies; broken equipment; and variability in responses as a function of occupancy, weather, and other variables.
3) **Observed DR variability** occurs as a result of the combination of unmodeled load variability and real DR variability.

*This paper is based on "Examining Uncertainty in Demand Response Baseline Models and Variability in Automated Responses to Dynamic Pricing" by J.L. Mathieu, D.S. Callaway, and S. Kiliccote which appears in the proceedings of the IEEE Conference on Decision and Control and European Control Conference (CDC-ECC) 2011. ©2011 IEEE. Johanna L. Mathieu is with the Department of Mechanical Engineering at the University of California at Berkeley; Duncan S. Callaway (corresponding author) is with the Energy and Resources Group at the University of California at Berkeley, 310 Barrows Hall, Berkeley, CA 94720-3050, USA; Sila Kiliccote is with the Lawrence Berkeley National Laboratory.

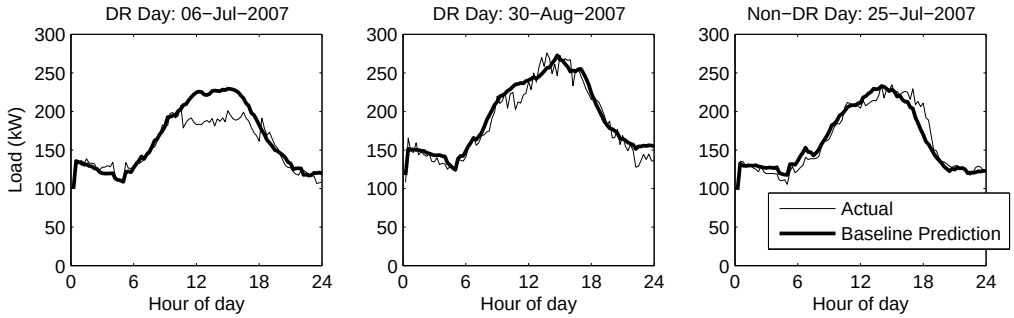

Fig. 1. Actual and baseline-predicted demand for an office building on three different days during the summer of 2007. The left and middle plots show data from DR event days (the difference between the actual and the baseline prediction is a combination of the response to the DR signal and model error), while the right plot shows data from a normal day (the difference between the actual and the baseline prediction is model error).

Fig. 1 illustrates the concepts of baseline model error and observed DR variability. In this figure, we plot the actual and baseline-predicted load for an office building on two DR days and one normal day. The left and middle plots show that responses to DR signals can seem variable – and may, in fact, be variable. The right plot demonstrates baseline model error.

The purpose of this paper is to understand the variability of C&I facility responses to DR events. The question is important for two reasons. First, in order to efficiently allocate generating resources, power system operators must predict how aggregations of facilities will respond on demand response days. If all observed DR variability resulted from unmodeled load variability, a power system operator could expect consistent DR behavior and would only need to deal with the usual amount of demand-side variability. However, if real DR variability is present, the DR program may create an extra burden of variability for the system operator to manage. This could require additional power system services (e.g., reserves). In extreme cases, real DR variability could result in significant deviations in grid frequency or expected power flow.

The second reason variability is important is because DR programs are evaluated on the basis of whether or not facilities (individually or in sum) appear to have reliable responses on DR days. M&V of utility DR programs, including those that do not use baselines for settlement, often include analyses of the DR performance (versus a baseline) of both individual facilities and aggregations of facilities [6]. Variability may affect the evaluation of the DR program and determinations about whether or not a facility is suitable for DR [8]. Moreover, observed DR variability in an individual facility affects how the facility perceives its own DR performance from event-to-event. A facility whose performance seems inconsistent from one event to another may be tempted to modify its DR strategy; however, the perceived inconsistency may have been caused by baseline model error.

Real DR variability is the most relevant measure for power system operators and DR program evaluators. However, real DR variability can only be estimated indirectly, by thoroughly characterizing unmodeled load variability and relating it to observed DR variability. Therefore, in this paper, we first compute the error associated with DR parameter estimates (e.g., demand shed estimates) for 38 C&I facilities that participated in an automated dynamic electricity pricing program in California. We then construct a variability metric that captures the relative importance of real DR variability versus unmodeled load variability, and compute this metric for all of the facilties. We find that most observed DR variability is the result of baseline model error.

A note on terminology: The DR community uses several different terms to denote the counterfactual power usage on DR days: baselines, predictions, and forecasts. In this paper, we use the term 'baseline predictions' to refer to ex-post estimates of counterfactual power usage computed with regression parameters (identified with historical demand/temperature data) and *actual* temperature data for the purpose of M&V. We reserve the term 'forecast' for ex-ante estimates computed with *forecasted* temperature data, which we do not discuss in this paper. We use the term 'DR parameter estimates' to refer to values, such as demand sheds, computed with actual demand data and baseline predictions. The DR community often refers to these values as 'DR calculations'; however, we prefer our terminology because it makes clear that the values are uncertain. The term 'DR parameter estimates' should not be confused with 'DR estimates,' engineering estimates of expected demand sheds.

The rest of this paper is organized as follows: In Sections II and III, we describe our data and baseline model. In Section IV, we explain our error analysis. Then, in Section V, we present our results and discussion with respect to baseline model error and DR variability. Lastly, in Section VI, we conclude.

II. Data

We use 15-minute interval whole building electric load data from 38 large C&I facilities (peak demand >200 kW) in California that participatied in Pacific Gas and Electric Company's (PG&E's) Automated Critical Peak Pricing (CPP) Program between 2006 and 2009. PG&E called CPP DR events on up to 12 summer business days (non-holiday, weekdays)

Table 1: Number of facilities by year and facility type.

	Office Buildings	Industrial Facilities	Retail Stores	Retail with PV*	Prisons & Jails	Museums	**TOTAL**
2006 (Zone 1)	3	0	1	0	0	0	**4**
2006 (Zone 2)	6	0	1	0	1	1	**9**
2007	7	1	3	1	1	1	**14**
2008	12	8	2	3	1	1	**27**
2009	17	8	1	3	3	1	**33**
TOTAL	**45**	**17**	**8**	**7**	**6**	**4**	**87**

*Retail stores with solar photovoltaics (PV).

per year when system-wide load was expected to be high, which, in California, usually occurs on hot summer days as a result of commercial and residential air conditioning. On DR days, electricity prices were raised to three times the normal price from 12 to 3 pm (moderate price period), and five times the normal price from 3 to 6 pm (high price period). In exchange for participating in the program, facilities paid lower energy prices on non-DR days. All 38 facilities used the OpenADR Communication Specification [5] to receive DR event notifications, which were provided by 3 pm the business day before the event. Each facility implemented a different set of pre-programmed DR strategies and executed the same strategies from event-to-event. Strategies included changes to the heating, ventilation, and air conditioning (HVAC) system, light dimming/switching, and industrial process shedding/shifting [9].

In 2006, DR events were called separately in two geographic zones: Zone 1 included San Francisco and the San Francisco Peninsula, while Zone 2 included the rest of PG&E's service territory. Nine DR events were called in Zone 1 and eleven in Zone 2. In both 2007 and 2009, twelve events were called, while in 2008 eleven events were called. Several facilities participated in only a portion of the DR events in a year. If we knew that a facility did not participate in a certain DR event, we did not analyze data from that DR day.

Facilities' demand profiles change year-to-year due to equipment upgrades, changes in usage patterns, etc. To reduce the chance of creating baseline models with data from before and after significant structural changes only one year worth of data were used to create each model. In total, we have 87 facility-years worth of data (Table 1), where a facility-year is defined as one year of data for one facility. Twelve facility-years of available data were not analyzed because of significant structural changes visible in the data.

Aggregate metrics were computed by summing power for all facilities for each year (and, in 2006, for each zone). We excluded facilities that did not participate in all of the DR events in a year and facility-years for which we were missing more than one week of data. In sum, nine facility-years were not included in the aggregate populations (hence the discrepancy in number of facilities between Tables 1 and 4). All aggregate results are computed from baseline models built with the aggregate data, not the aggregate output of individual baseline modes.

From the National Climatic Data Center [10], we acquired hourly outdoor air temperature data for each facility from the nearest weather station. Unfortunately, some of the temperature data are spotty. We linearly interpolated the data to assign an approximate temperature to every 15-minute interval, though when six or more hours of data are missing we do not interpolate. In some cases, when the data for a station were particularly spotty, we have filled the holes with data from another nearby station. Temperature data for the aggregate populations were generated by weighting and averaging data from the individual stations. Weights are determined by the number of facilities in the aggregate population associated with each station. For example, if N_1 facilities are associated with Station 1, which measured temperatures $T_1(t)$, and N_2 facilities are associated with Station 2, which measured temperature $T_2(t)$, and so on, then aggregate temperature, T_{agg}, at time step t is computed as follows:

$$T_{agg}(t) = \frac{\sum_{i=1}^{M} N_i T_i(t)}{\sum_{i=1}^{M} N_i} \quad , \tag{1}$$

where M is the total number of weather stations associated with facilities in the aggregate population.

III. Baseline Model & DR Parameters

Electric utilities generally use simple models to determine baseline electric load on DR days for financial settlement and/or M&V. Many of these models involve averaging the daily electric demand over several days (e.g., those with the highest energy usage) before the DR day [11], [12]. Unfortunately, baseline models built by averaging can be biased. Regression-based baseline models, which are less likely to suffer from bias, have long been used for M&V by the energy efficiency

community [13], [14], [15], [16] and are increasingly used for DR M&V [11], [12], [17], [18]. These models typically relate electric demand to weather and, sometimes, other relevant parameters. More sophisticated baseline modeling methods (e.g., neural networks) have been proposed, but are seldom used in practice.

We use the regression-based baseline model described in [19] because it performs similarly to or better than baseline models commonly used for DR M&V. Therefore, the magnitude of the error associated with this model is comparable to or less than that associated with common baseline models. Using better baseline models not only allows one to compute more accurate DR parameter estimates, but also allows one to better determine if a facility exhibits real variability in its responses to DR events.

A brief description of the baseline model is as follows: We expect demand to be a function of time-of-week. Regression coefficients, α_i, are assigned to each each 15-minute interval from Monday to Friday, t_i where $i = 1...480$. We also expect demand to be a piecewise linear and continuous function of outdoor air temperature, T, as described in [13], [15]. Observed temperatures are divided into six equal-sized temperature bins[1] and a regression coefficient, β_j where $j = 1...6$, is assigned to each bin. In each time step, each coefficient is multiplied by a temperature component $T_{c,j}$ computed from the current T using the following algorithm which ensures piecewise linearity and continuity [19]:

1) Let B_k for $k = 1...5$ be the interior bounds of the temperature intervals.
2) If $T > B_1$, then $T_{c,1} = B_1$. Otherwise, $T_{c,1} = T$ and $T_{c,m} = 0$ for m=2...6 and algorithm is ended.
3) For $n = 2...4$, if $T > B_n$, then $T_{c,n} = B_n - B_{n-1}$. Otherwise, $T_{c,n} = T - B_{n-1}$ and $T_{c,m} = 0$ for $m = (n+1)...6$ and algorithm is ended.
4) If $T > B_5$, then $T_{c,5} = B_5 - B_4$ and $T_{c,6} = T - B_5$.

We model the same temperature effect across all occupied mode hours (transitions between occupied and unoccupied are manually determined by looking at plots of average daily demand profiles on non-DR days). Estimated occupied mode demand, $\hat{D}_o$ is:

$$\hat{D}_o(t_i, T(t_i)) = \alpha_i + \sum_{j=1}^{6} \beta_j T_{c,j}(t_i) \quad . \tag{2}$$

We model a different temperature effect across all unoccupied mode hours. Since the facility often experiences a smaller range of temperatures during unoccupied mode (usually nighttime), we model the temperature effect as linear with only one regression coefficient, β_u, which is multiplied by outdoor air temperature T. Estimated unoccupied mode demand, $\hat{D}_u$ is:

$$\hat{D}_u(t_i, T(t_i)) = \alpha_i + \beta_u T(t_i) \quad . \tag{3}$$

Since all 2006-2009 DR days were called May 1 to Sept 30, baseline models were constructed with non-DR day demand data during the same period. We did not use data from holidays, weekends, or days that appeared to have had power outages (i.e. days when the minimum power use is less than a percentage of the average minimum daily power use during the summer) to build the baseline models.

The parameters α, β, and β_u are estimated with Ordinary Least Squares (OLS). We use the OLS estimator because, though it not 'best' (in a Gauss Markov sense) due to autocorrelation and heteroscedasticity (see Section IV), it still produces unbiased regression coefficients [20], [21]. However, the standard errors associated with the regression coefficients are underestimated, so we do not use them.

The parameter estimates and temperatures on DR days are then used to predict demand on DR days. Four DR parameters (Table 2), computed from the baseline predicted demand and the actual demand, are used to characterize changes in electricity use on DR days. These parameters are computed with 15-minute interval data and 15-minute interval predictions. For example, average demand shed is computed by averaging the 15-minute interval demand data during the event, averaging the 15-minute interval demand predictions during the event, and computing the difference of the averages. These parameters were first defined in [19]; however, here we define Daily Peak Demand and Daily Energy slightly differently: as absolutes, not percentages.

IV. Error Analysis

Most error analyses on regression-based baseline models use the standard errors associated with the regression coefficients [13], [16], [18]. However, these errors underestimate the true error due to a number of issues. First, the regression parameters are correlated. Specifically, time-of-week is correlated to temperature: the highest temperatures tend to occur in the afternoon and the lowest temperatures occur overnight. Second, the regression residuals are autocorrelated. In Fig. 2, we show autocorrelation functions (ACF) and partial autocorrelation functions (PACF) computed with regression residuals from two facility-years. In both cases, the residuals are lag 1 autocorrelated, which is the case for all facility-years. In some cases, we find higher order autocorrelation.

[1]Through trial and error, six bins were found to allow for enough change points and not cause over-fitting problems. This value is not optimized.

Table 2: DR parameter definitions, meanings, and importance.

Parameter	Definition	If this value is positive...	Importance
Average Demand Shed (kW)	Predicted minus actual average demand during the DR event.*	...the facility reduced power use during the event.	Key indicator for how well the facility performed.
Rebound (kW)	Actual minus predicted average demand in the hour after the DR event (6-7pm).	...the facility increased power use after the event.	Could affect a facility's demand charges; synchronized rebounds could create a new system-wide peak.
Daily Peak Demand (kW)	Actual minus predicted maximum demand on the DR day.**	...the facility had a higher demand peak than it would have if there was no DR event.	Could affect a facility's demand charges; will not affect the system-wide peak unless the individual peaks are synchronized.
Daily Energy (kWh)	Actual minus predicted total energy use on the DR day.	...the facility used more energy than it would have if there was no DR event.	Suggests whether energy shifting or shedding strategies predominate; implicated in understanding DR's effect on energy use and the environment, a research gap [23].

* The average demand shed is computed separately for each price period: 'Shed 1' refers to the moderate price period (12-3pm) and 'Shed 2' refers to the high price period (3-6pm).

** The actual and the baseline peak could happen at different times during the day.

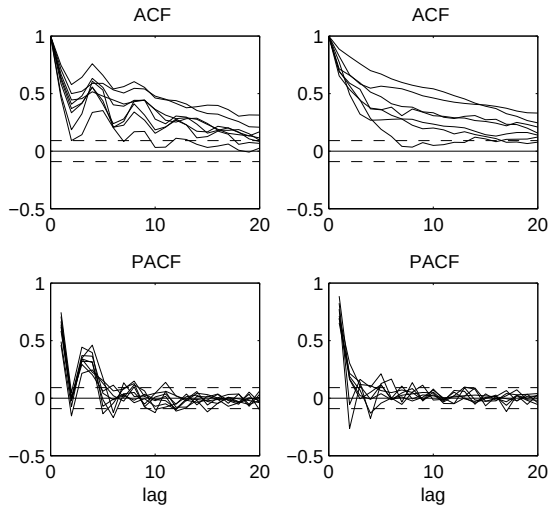

Fig. 2. ACF and PACF computed with the regression residuals from an office building (left) and a retail store (right) in 2008. Each line was created with data from a week (Mon-Fri) in which there were no DR days, holidays, or power outage days. Dashed lines show the 95% confidence interval ($\pm 2/\sqrt{n}$, where n is the number of data points in the data set).

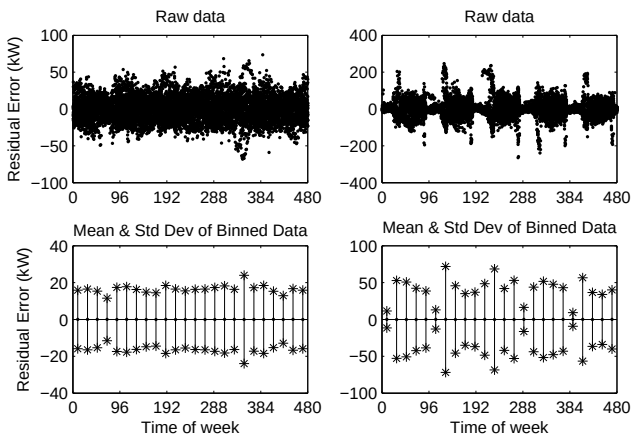

Fig. 3. Error versus Time-of-Week for an office building (left) and a retail store (right) in 2008.

Third, the regression residuals are heteroscedastic. Specifically, we find that the variance of the regression residuals (referred to as the 'error variance') is a function of time-of-week. For a typical commercial building, error variance tends to be lower at night and higher during the day when fluctuating occupancy affects loads. For some facilities, the error variance is high during transition periods (e.g., when the facility is being populated in the morning). Fig. 3 shows plots, created using (2) and (3), of error versus time-of-week. For the retail store, error is clearly a function of time-of-week, while for the office building, the effect is smaller. These results not only demonstrate heteroscedasticity, but also the importance of computing errors as a function of time-of-week. We have not computed error as a function of temperature or predicted demand because error does not seem to be a strong function of these variables.

These issues suggest that one should use caution in interpreting the standard errors associated with the baseline model regression coefficients. Fortunately, as we will explain in the method description below, we do not need to calculate this in order to calculate the error associated with DR parameter estimates.

A. Method

The goal of our error analysis is to determine the error associated with each DR parameter estimate for each facility-year and each aggregation of facilities. Other studies have used regression residuals to generate baseline model error estimates [17]; however, regression residuals are self-influenced: the model is built and tested on the same data set. Therefore, error estimates generated with regression residuals underestimate the true error.

To avoid self-influence, we use a resampling technique called 'Leave One Out Cross Validation' (LOOCV). LOOCV is a

type of K-fold cross validation, which involves randomly partitioning the data into K subsamples, reserving one subsample, building the model with data from the remaining subsamples, testing on the reserved subsample, and repeating this process for all K subsamples. The results for each subsample are combined resulting in an estimate of the prediction accuracy. In LOOCV, K is equal to the total number of observations, n. LOOCV is useful when n is small, though the technique is computationally intensive.

We treat the demand on each non-DR day as an observation. Therefore, n is equal to the number of non-DR days used to create the baseline prediction model ($\sim 90 - 95$ days per facility-year). We leave out one non-DR day, build the model with data from the rest of the non-DR days, predict the demand on the day that has been left out, compute the quantities associated with the DR parameters (e.g., average demand between 12 and 3 pm), compare the predictions to the actual quantities to generate an error observation, and repeat for each non-DR day. Since we consider error as a function of time-of-week, only residuals computed with data from Mondays are used to determine errors on Mondays, etc. Therefore, for each DR parameter for each day of week there are only $\sim 18 - 20$ error observations. It is difficult to determine the true error distribution with so few error observations. Therefore, we assume that the error observations are normally-distributed and report error estimates as one standard deviation of the error observations.

We do not recommend using this error analysis method on baseline models parameterized with DR day data (e.g., morning adjustments [12]). For those models, this method will underestimate true model error if power use outside of the DR period is affected by the DR signal, which is common, especially for facilities that pre-cool, rebound, or otherwise shift energy use to the morning or evening on DR days.

B. Other Sources of Error

Ideally, an error analysis should quantify all possible sources of error associated with an estimate. The methodology described above accounts for demand/temperature measurement error; error resulting from the fact that the weather stations are not co-located with the facilities; error resulting from temperature data interpolation; and unmodeled load variation on days similar to those used to build the baseline model. There are two sources of error we have not quantified: over-fitting and extrapolation. DR days are generally called on the hottest days of the summer which means that, in some cases, baseline predictions are made with temperatures: (1) higher than those on non-DR days, resulting in extrapolation error; and (2) experienced only a few times on non DR-days, resulting in over-fitting error. Over-fitting/extrapolation error could increase error variance and/or introduce bias in predictions made with high outdoor air temperatures.

For 26% of our DR day baseline predictions, the highest temperature on the DR day is greater than the highest temperature used to build the baseline model. We found that model error associated with extrapolated baseline predictions is comparable to that associated with non-extrapolated baseline predictions. Also, as described previously, error variance does not appear to be a strong function of temperature, even the highest temperatures used to build baseline models (i.e. temperatures for which we might expect possible over-fitting). Other baseline models, such as those that use fewer data to build the model, may be more susceptible to over-fitting/extrapolation error. Baseline models that model a load as a purely linear function of temperature may suffer from extrapolation error. If over-fitting/extrapolation error is present, the method presented in Section IV-A will underestimate DR parameter error. However, if error variance is not a function of temperature and consequently over-fitting/extrapolation introduces only bias, over-fitting/extrapolation error will not affect the accuracy of the DR parameter variability metrics (introduced in Section V-B).

V. Results & Discussion

A. DR Parameter Errors

The error analysis method presented in Section IV-A allows us to assign error estimates to DR parameter estimates. In Fig. 4, we show DR parameter and error estimates for all 2009 facility-years and the 2009 aggregate population. In most cases, the error estimates are large relative to the DR parameter estimates. For example, on average, across all years, the error associated with Shed 1 is approximately $\pm 120\%$ of the parameter values and the error associated with Shed 2 is approximately $\pm 180\%$ of the parameter values. In addition, observed DR parameter variability is often large. However, given the magnitude of the error estimates, we would expect some observed DR parameter variability.

This interpretation of Fig. 4 illustrates how including error estimates along with DR parameter estimates allows us to draw the right conclusions from the data. Without error estimates, it would be easy to classify a facility with observed shed variability as a variable shedder, and, therefore, conclude that such a facility is difficult to control. However, if the error associated with that facility's shed estimates is large, then it is possible that the response is actually consistent and we are simply unable to measure the exact response because of baseline model error.

There are several other things to learn from Fig. 4. Some facilities that shed power during DR events consume less energy on DR days, while some do not, meaning that they shift load outside of the DR period. We also find that the Daily Peak Demand is often biased low, because regression-based baseline models tend to under-predict maximum values (i.e. outliers). The aggregate population results demonstrate that DR works: the aggregated facilities shed power during DR events and reduce the peak demand on DR days, despite the fact that individual facilities may become peakier. Also, on average,

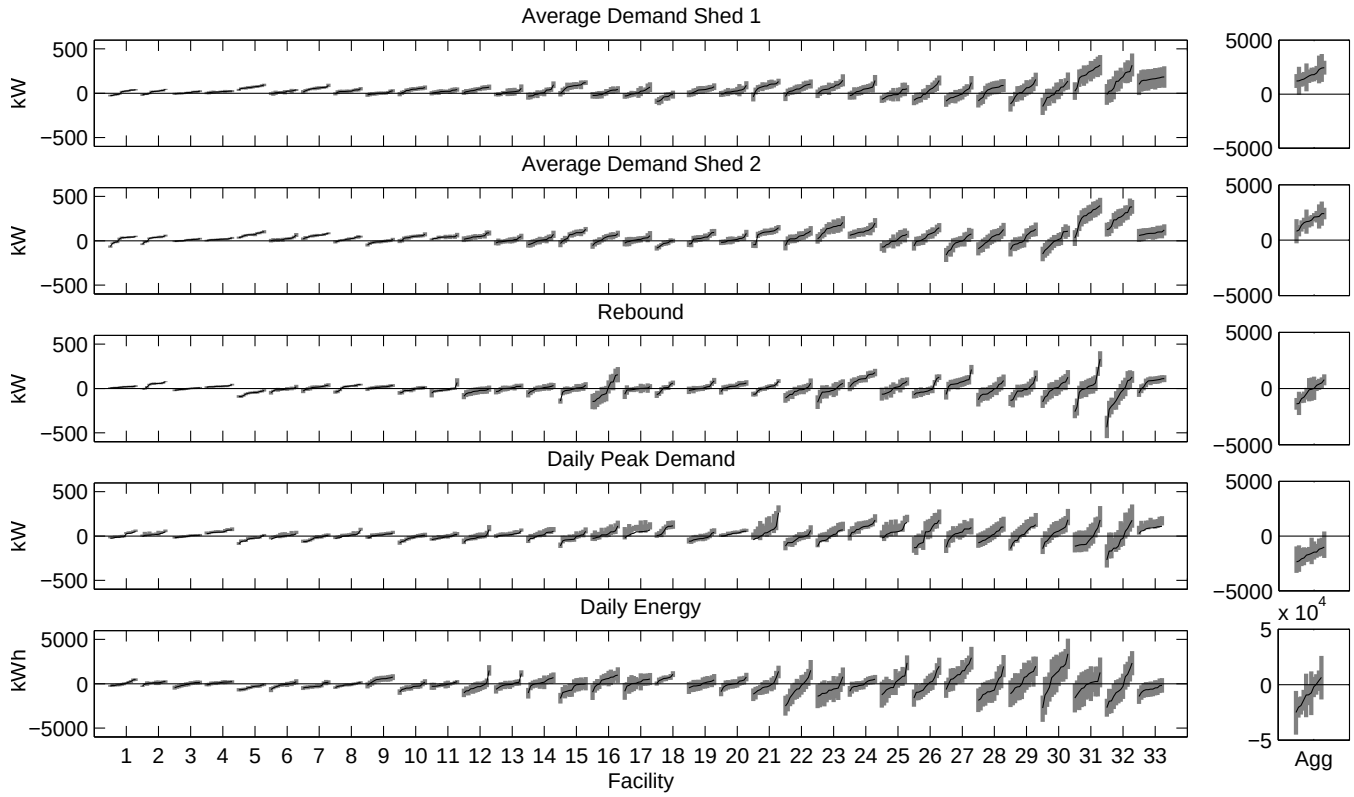

Fig. 4. DR parameter estimates (black) and error estimates (grey) for all 2009 facility-years. Facilities are arranged in order of smallest to largest mean error for Average Demand Shed 1. For each facility and each DR parameter, parameters are arranged in order of smallest to largest. Results for the 2009 aggregate population are shown on the right.

the aggregated facilities exhibit almost no rebound and save some energy on DR days, indicating that there is some net curtailment–the facilities do not simply shift all load outside of the DR period.

We do not discuss the statistical significance of the DR parameter estimates because the error estimates are not confidence intervals. Since a facility's DR behavior from one DR event to the next is not independent, Bayesian techniques should be used to not only determine appropriate confidence intervals, but also pinpoint DR parameter estimates. This would involve pooling information across DR events (i.e. using knowledge about a facility's behavior during one DR event to help us predict its behavior during another DR event). We do not tackle this here because we are interested in using the error estimates to assess DR parameter variability, not statistical significance.

B. DR Parameter Variability

Observed DR parameter variability has two possible sources: unmodeled load variability and real parameter variation. For example, consider the Average Demand Shed. We generally observe shed variability from one DR event to the next. We would like to know if observed shed variability is a result of real shed variability (i.e. a facility curtails a different amount from event-to-event) or if it results from unmodeled load variability (i.e. baseline model error). If observed shed variability results exclusively from unmodeled load variability, then we can expect consistent responses and the system operator need only deal with the usual level of demand-side variability. If real shed variability exists, the system operator may require additional reserves to deal with more demand-side variability than usual.

In Section V-B.1, we derive a metric, the Average Demand Shed Variability Metric (SVM), to discern between unmodeled load variability and real DR parameter variation. Similar derivations yield metrics for each DR parameter: the Rebound Variability Metric (RVM), Daily Peak Demand Variability Metric (PVM), and Daily Energy Variability Metric (EVM). In Sections V-B.2 and V-B.3, we present DR parameter variability metric results for the individual facility-years and the aggregate populations, respectively.

1) SVM Derivation: On a DR day, the *Observed Load* (OL) is equal to the *Real Baseline Load* (RBL) minus the *Real Shed* (RS):

$$OL = RBL - RS \quad . \tag{4}$$

Neither the real baseline load nor the real shed can be measured. The real baseline load is estimated with the *Predicted Baseline Load* (PBL). The difference between the real baseline load and the predicted baseline load is the *Unmodeled Load* (UL):

$$UL = RBL - PBL \quad . \tag{5}$$

To compute the *Observed Shed* (OS), the predicted baseline load is subtracted from the observed load:

$$OS = OL - PBL = UL - RS \quad . \tag{6}$$

Our goal is to determine the variance of the real shed. Therefore, we take the variance of (6), which results in:

$$\text{Var}(OS) = \text{Var}(UL) + \text{Var}(RS) - 2\text{Cov}(UL, RS) \quad . \tag{7}$$

We can estimate Var(OS) by taking the variance of the $9 - 12$ observed sheds and Var(UL) by taking the variance of the ~ 95 error observations (since DR events can occur on any weekday, error observations are used without regard to day-of-week). Therefore, we define the shed variability metric (SVM) as:

$$\begin{aligned} \text{SVM}: &= \text{Var}(OS) - \text{Var}(UL) \\ &= \text{Var}(RS) - 2\text{Cov}(UL, RS) \quad . \end{aligned} \tag{8}$$

While the SVM does not tell us the exact value of Var(RS) due to the complicating covariance term, it does tell us if real shed variability likely exists or not. Also, since Var$(RS) \geq 0$, the SVM may tell us something about the sign of the covariance term. If the covariance term is positive, then as unmodeled load increases, real shed increases. This could occur when the equipment that drives the unmodeled load is also the equipment that is curtailed. Alternatively, if the covariance term is negative, then as unmodeled load increases, real shed decreases. This could occur when load is higher than predicted, electricity consuming services are in high demand, and occupants/building operators override automated DR strategies; or when load is higher than predicted, the HVAC system is operating at or beyond its maximum capability, and consequently a reduction in HVAC setpoint has a limited effect.

2) Individual Facility-years: To compare facilities by SVM, we normalize the measurements of the unmodeled load and the observed shed such that Var(UL) $= 1$. Therefore, the minimum value of SVM is -1 (i.e. when Var(OS) $= 0$). Each DR parameter variability metric is normalized similarly.

Histograms showing DR parameter variability metrics for the 87 facility-years are shown in Fig. 5. To understand what these histograms tell us about real parameter variability, we can compare them to distributions generated for the case when real parameter variability is zero. If real parameter variability were zero, the covariance term would also be zero, resulting in a DR parameter variability metric of zero. However, we are unable to compute the 'true' values of the DR parameter variability metrics because we can only estimate observed parameter variance from ~ 11 observations. Assuming that the observations are normally-distributed, we would expect the distribution of observed parameter variances to follow a scaled χ^2 distribution with $N - 1$ degrees of freedom [22]:

$$\frac{(N-1)x}{\sigma^2} \sim \chi^2_{N-1} \quad , \tag{9}$$

where x is the sample variance, N is the number of observations, and σ^2 is the true variance. Therefore, the expected variability metric distributions for the case when real variability is zero is that given in (9), shifted left by 1 (resulting from the subtraction of Var$(UL) = 1$ in (8)). These distributions (for $N = 11$) are plotted in Fig. 5. One caveat associated with these results is that we have assumed that we know the 'true' value of Var(UL), though, in reality, it is an estimate (computed from ~ 95 observations). When we normalize the measurements of the unmodeled load and the observed shed such that Var(UL)=1, any error in our estimate of Var(UL) will affect our estimate of Var(OS), which, in turn, affects our estimate of the SVM.

If none of the facility-years exhibited real parameter variability then we would expect only 5% of facilities to fall outside of the 95% confidence bounds. However, for each parameter, we find that substantially more than 5% of the facility-years fall outside of the bounds (Table 3). This implies that some facility-years exhibit real parameter variability. Facilities with disproportionally positive variability metrics likely exhibit real parameter variability. Facilities with disproportionally negative variability metrics likely exhibit positive covariance and, subsequently, real parameter variability. For the remainder of the facility-years, any observed parameter variability may simply result from model error and sampling.

Through simulation we find that, in order to achieve the distributions shown in Fig. 5, it is likely that a number of facility-years have large real parameter variability, while the majority of facility-years have little to no real parameter variability. Also, it is likely that for the vast majority of facility-years the covariance term is positive which implies that as unmodeled load increases, real shed increases, which is consistent with intuition. Additionally, we find that all combinations of the variability metrics are positively correlated. Several combinations have a population correlation coefficient, $\rho_{x,y}$, greater than 0.5: SVM1 and SVM2 ($\rho_{x,y} = 0.76$); SVM1 and PVM ($\rho_{x,y} = 0.59$); and SVM2 and PVM ($\rho_{x,y} = 0.56$).

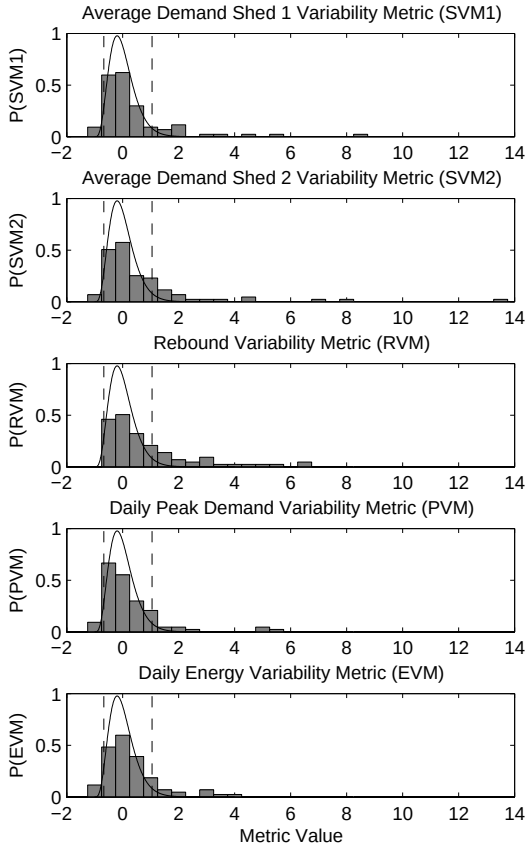

Fig. 5. Histograms showing DR parameter variability metrics for the 87 facility-years. Solid lines show the expected distributions if real parameter variability were zero and $N = 11$ (dashed lines show the 95% confidence interval). Disproportionally positive variability metrics result from real parameter variability. Disproportionally negative variability metrics result from negative covariance and, subsequently, real parameter variability.

Table 3: Number and percentage of facility-years with variability metrics inside and outside the 95% confidence bounds.

Metric	Inside Bounds	Outside Bounds		
		Below	Above	TOTAL
SVM1	65 (75%)	8 (9%)	14 (16%)	22 (25%)
SVM2	62 (71%)	4 (5%)	21 (24%)	25 (29%)
RVM	62 (71%)	2 (2%)	23 (26%)	25 (29%)
PVM	71 (82%)	6 (7%)	10 (11%)	16 (18%)
EVM	69 (79%)	6 (7%)	12 (14%)	18 (21%)

* Percentages do not always add properly due to rounding.

Table 4: DR parameter variability metrics computed for the aggregate populations. Bold values indicate p-values $\leq$0.05.

Year	Facilities (Peak*)	Shed 1		Shed 2		Rebound		Daily Peak Demand		Daily Energy	
		SVM1	p-value	SVM2	p-value	RVM	p-value	PVM	p-value	EVM	p-value
2006 (Zone 1)	4 (2.7 MW)	**-0.819**	**(0.01)**	-0.269	(0.67)	0.077	(0.75)	-0.386	(0.47)	**-0.737**	**(0.04)**
2006 (Zone 2)	8 (8.4 MW)	**3.039**	**(<0.01)**	**3.399**	**(<0.01)**	**1.044**	**(0.05)**	**1.131**	**(0.04)**	**4.578**	**(<0.01)**
2007	13 (11.7 MW)	0.579	(0.21)	-0.117	(0.90)	-0.454	(0.32)	-0.531	(0.24)	-0.210	(0.78)
2008	21 (14.6 MW)	-0.210	(0.72)	-0.142	(0.86)	**1.295**	**(0.02)**	-0.217	(0.71)	0.163	(0.62)
2009	32 (26.9 MW)	**-0.696**	**(0.03)**	-0.331	(0.46)	0.304	(0.43)	**-0.702**	**(0.04)**	-0.227	(0.69)

*Peak demand computed for May 1 - Sept 30.

The Federal Energy Regulatory Commission's (FERC) has called for better understanding of responses to dynamic prices as a function of customer type [23], so we attempted to disaggregate parameter variability results by facility attributes including facility type, HVAC system type, DR strategy, and shed size. Results were inconclusive because of the small number of facility-years in the data set. It was particularly difficult to disaggregate the facilities by DR strategy because many facilities use more than one strategy. Therefore, we were unable to determine what kinds of facilities have more or less variable DR parameters. In an effort to do this, we are in the process of acquiring a larger data set.

3) Aggregate Populations: DR parameter variability metrics for each aggregation of facilities are shown in Table 4. For each variability metric, we have computed the two-sided p-value under the null hypothesis that there is no real parameter variability. Therefore, real parameter variability likely exists when p-values are small. Surprisingly, the aggregate populations exhibit a wide range of variability metrics, similar to that seen for the individual facility-years. We would expect more real DR parameter variability in smaller aggregate populations. For example, in 2006 Zone 2 (8 facilities), we find likely real variability in each DR parameter. However, we also find likely real variability in both the Average Demand Shed 1 and the Daily Peak Demand in 2009 (32 facilities). Real variability in the aggregate could result from unmodeled correlation across facilities and/or large variable facilities dominating the aggregate results.

VI. Conclusions

We have developed a method to determine the error associated with DR parameter estimates. We find that this error is often large and so DR parameter estimates reported without error estimates may be misleading. For example, we may classify a steady shedder as a variable shedder and, therefore, judge the facility to be poorly controlled when, in fact, baseline model error simply prevents us from measuring consistent sheds. Since DR parameter estimates have error, all calculations derived with these estimates, including cost effectiveness estimates, also have error. Future research should explore the degree to which DR parameter error affects cost/benefit analyses on DR programs and technologies.

Observed DR variability in both individual facilities and aggregations of facilities affects the perception of persistent and reliable demand sheds. However, we find that observed DR variability is driven, in large part, by baseline model error, not real DR variability. For most facilities, observed DR variability can likely be explained by baseline model error alone; however, a number of facilities likely exhibit high real DR variability. In addition, most facilities exhibit a positive correlation between unmodeled load and real shed, which implies that the equipment that drives baseline model error is the equipment that is curtailed during DR events. We were unable to discern any relationship between response variability and facility attributes, possibly due to the small number of facilities.

Variability metrics computed for the aggregate populations show that in some cases the aggregate population likely exhibits real DR variability, which has implications for utilities with respect to M&V of DR programs. It also has implications for the system operator. If the aggregate response is not consistent, the system operator may have to deal with more demand-side variability than exists on non-DR days and, therefore, will need to procure more power systems services. In extreme cases, DR variability could result in significant deviations in grid frequency or expected power flow. More research is needed to understand DR variability in aggregate populations composed of facilities executing manual DR strategies, as they may exhibit even more variability than populations composed of facilities executing automated strategies.

The DR signal considered here, a critical peak pricing signal, is open-loop since the prices do not change in response to changes in load (though the signal is often implemented in individual facilities as closed-loop indoor air temperature control). Our results would be different if a closed-loop DR signal were used. Specifically, we would expect less DR variability, which could mitigate some of the issues we have described. This is an important subject of future research.

ACKNOWLEDGMENTS

We thank Phillip Price, Mary Ann Piette, and Ashok Gadgil for great advice and feedback. We also thank PG&E Company for the electric load data. Johanna Mathieu was funded by a UC Berkeley Chancellor's Fellowship. Some of this work was conducted at the Lawrence Berkeley National Laboratory under U.S. Department of Energy Contract No. DE-AC02-05CH11231. Sila Kiliccote was funded by the California Energy Commission (CEC) under Contract No. 500-03-026.

REFERENCES

[1] Y. Sherif and S. Zahir, "Communication systems for load management," *IEEE Transactions on Power Apparatus and Systems*, vol. PAS-104, no. 12, pp. 3329–3337, December 1985.

[2] J. Eto, J. Nelson-Hoffman, C. Torres, S. Hirth, B. Yinger, J. Kueck, B. Kirby, C. Bernier, R. Wright, A. Barat, and D. Watson, "Demand response spinning reserve demonstration," Lawrence Berkeley National Laboratory LBNL-62761, 2007.

[3] ERCOT Wholesale Market Subcommittee Demand-Side Working Group, "Load participation in the ercot nodal market," Electricity Reliability Council of Texas, Tech. Rep., June 2010.

[4] D. Callaway and I. Hiskens, "Achieving controllability of electric loads," *Proceedings of the IEEE*, vol. 99, no. 1, pp. 184–199, 2011.

[5] M. Piette, G. Ghatikar, S. Kiliccote, E. Koch, D. Hennage, P. Palensky, and C. McParland, "Open automated demand response communications specification (Version 1.0)," California Energy Commission (PIER Program) CEC-500-2009-063, 2009.

[6] G. Wikler, I. Bran, S. Prijyandonda, S. Yoshida, K. Smith, M. Piette, S. Kiliccote, G. Ghatikar, D. Hennage, and C. Thomas, "Pacific gas & electric company 2007 Auto-DR program: Task 13 deliverable: Auto-DR assessment study," Pacific Gas and Electric Company, Tech. Rep., Dec. 2007. [Online]. Available: http://drrc.lbl.gov/pubs/pge-auto-dr-assessment-study.pdf

[7] D. Hansen, "2009 load impact evaluation and cost effectiveness tests of california statewide automated demand response programs," Christensen Associates Energy Consulting CALMAC Study ID: SDG0246.01, Sept. 2010.

[8] C. A. E. Consulting, "Highly volatile-load customer study for southern california edison, pacific gas and electric company, and san diego gas & electric, Christensen Associates Energy Consulting," Oct. 2010.

[9] N. Motegi, M. Piette, D. Watson, S. Kiliccote, and P. Xu, "Introduction to commercial building control strategies and techniques for demand response," Lawrence Berkeley National Laboratory LBNL-59975, 2007.

[10] NOAA, "National climatic data center," http://www7.ncdc.noaa.gov/CDO/dataproduct, 2009. [Online]. Available: http://www7.ncdc.noaa.gov/CDO/dataproduct

[11] M. Goldberg and G. Agnew, "Protocol development for demand response calculation- findings and recommendations," California Energy Commission CEC-400-02-017F, 2003.

[12] K. Coughlin, M. Piette, C. Goldman, and S. Kiliccote, "Statistical analysis of baseline load models for non-residential buildings," *Energy and Buildings*, vol. 41, no. 4, p. 374, 2009.

[13] M. Fels, "PRISM: An introduction," *Energy and Buildings*, vol. 9, p. 5, 1986.

[14] S. Katipamula, T. Reddy, and D. Claridge, "Multivariate regression modeling," *Journal of Solar Energy Engineering*, vol. 120, pp. 177–184, 1998.

[15] J. Kissock, T. Reddy, and D. Claridge, "Ambient-temperature regression analysis for estimating retrofit savings in commercial buildings," *Journal of Solar Energy Engineering*, vol. 120, p. 168, 1998.

[16] J. Kissock and C. Eger, "Measuring industrial energy savings," *Applied Energy*, vol. 85, no. 5, pp. 347–361, 2008.

[17] S. Kiliccote, M. Piette, and D. Hansen, "Advanced controls and communications for demand response and energy efficiency in commercial buildings," in *Proceedings of 2006 Carnegie Mellon Conference in Electric Power Systems*, Pittsburgh, PA, 2006.

[18] S. Braithwait, D. Hansen, and J. Reaser, "2009 load impact evaluation of California statewide critical-peak pricing rates for non-residential customers: Ex post and ex ante report," Christensen Associates Energy Consulting CALMAC Study ID: SDG0244.01, 2010.

[19] J. Mathieu, P. Price, S. Kiliccote, and M. Piette, "Quantifying changes in building electricity use with application to demand response," *IEEE Transactions on Smart Grid*, p. *in press*, 2011.

[20] T. Reddy, J. Kissock, and D. Ruch, "Uncertainty in baseline regression modeling and in determination of retrofit savings," *Journal of Solar Energy Engineering*, vol. 120, p. 185, 1998.

[21] D. Ruch, J. Kissock, and T. Reddy, "Model identification and prediction uncertainty of linear building energy use models with autocorrelated residuals," in *Proceedings of 1993 ASME International Solar Energy Conference*, 1993, p. 465.

[22] G. Vining, *Statistical Methods for Engineers*. Duxbury Press, 1998.

[23] FERC, "National action plan on demand response," Tech. Rep., 2010. [Online]. Available: http://www.ferc.gov/legal/staff-reports/06-17-10-demand-response.pdf