Contents lists available at ScienceDirect

Energy and Buildings

journal homepage: www.elsevier.com/locate/enbuild

Statistical analysis of baseline load models for non-residential buildings

Katie Coughlin [*], Mary Ann Piette, Charles Goldman, Sila Kiliccote

Demand Response Research Center, Lawrence Berkeley National Laboratory, One Cyclotron Road, Berkeley, CA 94705, United States

ARTICLE INFO

Article history:
Received 30 April 2008
Received in revised form 10 November 2008
Accepted 10 November 2008

Keywords:
Demand response
Baseline load profile
Impacts estimation

ABSTRACT

Policymakers are encouraging the development of standardized and consistent methods to quantify the electric load impacts of demand response programs. For load impacts, an essential part of the analysis is the estimation of the baseline load profile. In this paper, we present a statistical evaluation of the performance of several different models used to calculate baselines for commercial buildings participating in a demand response program in California. In our approach, we use the model to estimate baseline loads for a large set of proxy event days for which the actual load data are also available. Measures of the accuracy and bias of different models, the importance of weather effects, and the effect of applying morning adjustment factors (which use data from the day of the event to adjust the estimated baseline) are presented. Our results suggest that (1) the accuracy of baseline load models can be improved substantially by applying a morning adjustment, (2) the characterization of building loads by variability and weather sensitivity is a useful indicator of which types of baseline models will perform well, and (3) models that incorporate temperature either improve the accuracy of the model fit or do not change it.

Published by Elsevier B.V.

1. Introduction

Both federal and state policymakers are encouraging the development of standardized and consistent approaches to quantify the load impacts of demand response programs. In their report to Congress on Demand Response and Advanced Metering, the Federal Energy Regulatory Commission identified the need for consistent and accurate measurement and verification of demand response as a key regulatory issue [1,2]. The California Public Utility Commission is currently overseeing a regulatory process to develop methods to estimate the load impacts of demand response (DR) programs, which will help to measure their cost-effectiveness, assist in resource planning and long-term forecasting exercises, and allow the California Independent System Operator to more effectively utilize DR as a resource.

Policymakers are concerned that the methods used to estimate load reductions lead to fair and accurate compensation for DR program participants, and provide useful information to resource planners and system operators who wish to incorporate demand-side programs into the resource mix. Challenges in estimating load impacts include the diversity of customer loads and curtailment strategies, the heterogeneity in types of demand response programs and dynamic pricing tariffs, and variability in event characteristics such as timing, duration, and location [2]. Given the variability in the loads being modeled and the diversity of potential model applications, it is useful to have a general framework for evaluating the performance of different load impact estimation methods.

This paper describes a new statistical analysis of the performance of different models used to calculate the baseline electric load for buildings participating in an event-driven demand response program. During a demand response event, a variety of adjustments may be made to building operation, with the goal of reducing the building peak electric load. In order to determine the actual peak load reduction, an estimate of what the load would have been without any DR actions is needed. This is referred to as the *baseline load profile* (or baseline) and is key to accurately assessing the load impacts from certain types of demand response programs, particularly those that pay directly for load reductions [3].

The sample of buildings included in this study consists of participants in a California Automated Demand Response pilot [4,5] sponsored by the California Energy Commission through the Demand Response Research Center. The sample covers a wide range of building sizes and activities, including commercial (office and retail), institutional (schools, government) and a few industrial facilities including a bakery, electronics manufacturing and mixed-use [3]. For each of these facilities, a set of demand response strategies has been developed and field tested in collaboration with building owners and managers. Examples of strategies include dimming or turning off non-critical lights, changing thermostat set-points, or turning off non-critical equipment. The

[*] Corresponding author.
E-mail address: kcoughlin@lbl.gov (K. Coughlin).

0378-7788/$ – see front matter. Published by Elsevier B.V.
doi:10.1016/j.enbuild.2008.11.002

site characteristics, selection and implementation of control strategies, and results of field tests have been extensively documented in reports available at the Demand Response Research Center web site [3–5,14]. The appendices of references [4,5,14] provide summary information for the sites included in this study.

In the cases examined here DR events are only called on normal working days, during the period from 12 pm to 6 pm. These events are called during times of system stress, which are typically related to weather. For California, DR is used in the summer to deal with high peak loads on weekdays, which are often driven by space cooling in buildings. Given the correlation between temperature and increased building energy use for space conditioning, non-weather corrected models may under-predict the baseline and therefore systematically underestimate the response. This can be true even for buildings with large non-weather responsive loads, if the weather-dependent load is significant relative to the estimated load reduction. On the other hand, many customers, load aggregators and program administrators have a strong preference for simpler calculation methods with limited data requirements. It is useful therefore to establish how much quantitative improvement is gained by introducing more complicated calculation methods.

1.1. Analytical approach

Our main objective in this work is to provide a statistically valid evaluation of how well different baseline models perform, and to relate the performance to more general building characteristics. To do so, we need to define both the sampling procedure and the evaluation metrics. Building loads always have a random component, so the baseline estimation problem is inherently statistical in the sense that to properly assess the performance of a method, a sufficiently large sample of applications must be considered. Because our building sample is small, to develop a large enough data set, we define a set of *proxy event days* (days on which no curtailment occurs and the load is known, but which are similar in terms of weather to actual event days). For these days, we use the historical data and the model to predict the load, and compare the prediction to the actual load for that day. If the proxy event set is large enough, we can evaluate each model for each site separately. We focus on metrics that quantify the bias and the accuracy of the model at the building level.

In this study we evaluate seven models, for a sample of 32 sites in California incorporating 33 separately metered facilities. In some cases the meter may include electricity use for multiple buildings at one location. The models considered here can be broadly categorized into *averaging methods* and *explicit weather models*. For each baseline model, we tested two implementations: one without and one with a morning adjustment (which incorporates site usage data from the morning of the DR event prior to load curtailment). For each site, 15-min electric interval load data are available through the web-based customer energy metering site maintained by Pacific Gas and Electric.

1.2. Prior work

Several recent studies have reviewed and analyzed alternative methods for calculating the load reduction impacts of demand response programs [6–8]. The most extensive review of baseline modeling methods is provided in the Kema 2003 study *Protocol Development for Demand Response Calculation—Findings and Recommendations* [6]. This study examined a number of methods in use by utilities and electric system operators across the country, and evaluated them in terms of accuracy and bias. As noted there, a baseline calculation method is defined by specifying three component steps: a set of data selection criteria, an estimation method, and an adjustment method. The estimation uses load data for a period prior to the event day to predict the baseline load profile during the event period, while the adjustment uses data from the event day, before the beginning of the curtailment period, to align and shift the predicted load profile by some constant factor to account for characteristics that may affect load on the day of the event.

The Kema 2003 report [6] included only three accounts from California in their total sample of 646 accounts. Their sample is dominated by data from the eastern U. S. Given significant climatic and demographic variation across the country, with corresponding differences in building practices, occupancy, etc., it is unclear how well results will generalize across different regions. In particular, the Kema study found that explicitly weather–dependent models did not generally outperform models that did not include weather. One of the goals of this work is to determine whether this hypothesis also holds true for California.

Quantum Consulting conducted an analysis of baseline estimation methods as part of its broader evaluation of California's 2004 DR programs for industrial and commercial customers [7]. Their study used billing data for a sample of 450 customers eligible for the DR programs with peak demand ranging from 200 kilowatts to greater than 5 megawatts. Eight proxy event days were selected for each utility from the period July 1, 2003 to August 31, 2003.

In developing the statistical sample of test profiles, these studies used a large number of accounts, but a relatively small number of calendar days. Our statistical approach is different, using a much larger selection of proxy event days. This allows us to create a statistical picture for each building, and evaluate whether different methods perform equally well for different building types. We also present a new method for estimating the degree of weather-sensitivity of a building, and different diagnostics to quantify the predictive accuracy of the baseline model, and the estimated peak load savings values that are used in bill settlement. The metric used for measuring bias in the model results, and some of the baseline models tested, are similar to the earlier studies [6–8].

The remainder of the paper is organized as follows: In Section 2 we present an overview of the technical steps involved in preparing the data sets, defining the sample of proxy event days, running the models and developing the diagnostics. In Section 3 we describe the metrics we use to quantify weather sensitivity and load variability. In Section 4 we define each of the methods investigated in this paper, with results presented in Section 5. In Section 6 we present our conclusions and some suggestions for future work.

2. Data processing and evaluation metrics

In this section we describe the preparation of the data, the mechanics of implementing different models, and the diagnostic metrics used in this report.

2.1. Data sources

For each site examined in this project, the available 15-min electric interval meter data are converted to hourly time series by averaging the values in each hour. We use data from May to October of 2005 and 2006 to define the sample days and test the methods. Only the warm-weather months are included here, as these are the periods when (to date) events have been called in California's DR Programs. The amount of data available depends on how long the account has participated in the program (in some cases interval meters were installed because the site was willing to go onto a DR program), and whether there is any missing data during the sample period.

The explicit weather models require hourly temperature data for each site. The data were obtained by assigning each site to a weather station that is currently active and maintained by either a state or a federal agency [3–5]. The sites were chosen from those maintained by the National Oceanic and Atmospheric Administration (available by subscription) [9] or by the California Irrigation Management Information System [10], which is a program of the state Department of Water Resources. Only outdoor dry bulb air temperature data were used in developing the weather-dependent models.

2.2. Proxy event days

The goal of using proxy event days is to have a large sample set for which (i) the actual loads are known and (ii) the days are similar in some sense to the historical DR event days called by the California Independent System Operator in 2005 and 2006. Before selecting the proxy set, we first need to define the set of what we call *admissible days*, which is the set of days that can be used as input to the baseline load model calculations. We define admissible days as normal working days, i.e., eliminating weekends, holidays and past curtailment events, which follows standard procedures [6].

The proxy event days are selected as a subset of the admissible days. DR events are typically called on the hottest days, and can be called independently in each of several climate zones defined for California (all the sites available for this study are located in either zone 1 or zone 2 [3]). To define the weather characteristics associated with an event day, we first construct a spatially averaged zonal hourly temperature time series, using a simple average over the weather stations located in the zone. The hourly zonal temperatures are then used to construct three daily metrics: the maximum daily temperature, the average daily temperature, and the daily cooling degree hours using 65 F as the base temperature.

Sorting the weather data on the value of the daily metric provides a list of the hottest to coolest days in the sample period. We defined the proxy event days as the top 25 percent of the admissible days sorted in this manner. The three metrics give consistent results for the hottest days, but select slightly different samples. A little over 3/4 of the historical event days in each year are included in the top 25 percent selected. The results presented here use the sample associated with cooling-degree hours. For each building, a proxy event day is included in the analysis only if there is sufficient load data for that day. Hence, the proxy event sets vary somewhat from building to building. On average, this procedure leads to about 60 proxy event days for each site.

2.3. Model runs and diagnostics

Model results are calculated for all the admissible days, but diagnostics are calculated only for the set of proxy event days. For each model and each building site, the baseline load estimate for each hour from 9 am to 6 pm is calculated. While the event period is limited to 12 pm–6 pm, the adjustment factors require model and actual data from the early morning period. Using d and h to label the day and hour, respectively, we define the predicted load as pl(d,h), the actual load as al(d,h), and the adjustment factor for day d as c(d). The absolute difference between the actual load and the predicted load is defined as x(d,h) = al(d,h) − pl(d,h), and the relative difference between the actual load and the predicted load (or percent error) e(d,h) is defined as the ratio x(d,h)/al(d,h). For a given combination of a model and a site we calculate x(d,h) and e(d,h) for each proxy event day and each hour in the event period, which gives about 360 observations.

Utilities or system operators often settle payments for performance during DR events based on the average hourly load reduction during the hours of the event. It is therefore also useful to compare the prediction of the average hourly load to the actual value. We define A(d) and P(d) as the actual and predicted hourly load averaged over the event period, respectively. The absolute difference between the actual and predicted average event-period hourly load is X(d), and the percent difference is the ratio E(d) = X(d)/A(d). This notation is summarized in Table 1.

2.3.1. Adjustment factors

We evaluate each model both with and without a *morning adjustment* factor applied. We use a multiplicative factor defined as the ratio of the actual to the predicted load in the two hours prior to the event period:

$$c(\mathrm{d}) = \frac{[\mathrm{al}(\mathrm{d}, \mathrm{h} = 10) + \mathrm{al}(\mathrm{d}, \mathrm{h} = 11)]}{[\mathrm{pl}(\mathrm{d}, \mathrm{h} = 10) + \mathrm{pl}(\mathrm{d}, \mathrm{h} = 11)]} \quad (1)$$

To adjust the output of the baseline model, we multiply the predicted value in each hour by the daily adjustment factor. This factor essentially scales the customer's baseline from admissible days to the customer's operating level on the actual day of a DR event. We also tested an alternative adjustment approach that used the two hours preceding the event to define an additive, rather than multiplicative, correction factor. In our sample, there is no significant difference in the results.

Deciding on the period to use for the adjustment factor can be problematic for DR programs or tariffs where the event is announced on prior days, as there may be some concern about customers "gaming" their baseline by intentionally increasing consumption during the hours just prior to the event [6]. This is not a concern for the proxy event days used in this analysis, however it may lead to some differences in how well the model performs for proxy as compared to real event days.

2.3.2. Diagnostic measures

For each model, both with and without adjustment, and each site, we calculate the set of absolute and percentage errors x(d,h) and e(d,h). Our evaluation of the performance of a model is based on the statistical properties of these errors. To measure any bias in the model, we calculate the median of the distribution of errors. If the method is unbiased the median will be zero. If the median is positive (negative) it means that the model has a tendency to predict values smaller (larger) than the actual values. To quantify the accuracy of the

Table 1
List of symbols and labels used in the text.

Symbol	Definition
h	Hour index
d	Admissible day index
c(d)	Morning adjustment factor
pl(d,h)	Predicted load for day d and hour h
al(d,h)	Actual load for day d and hour h
x(d,h)	Absolute difference between actual and predicted load: x(d,h) = al(d,h) − pl(dh)
e(d,h)	Relative difference between actual and predicted load: e(d,h) = x(d,h)/al(d,h)
P(d)	Predicted load averaged over the event period
A(d)	Actual load averaged over the event period
X(d)	Absolute difference: X(d) = A(d) − P(d)
E(d)	Relative difference: E(d) = X(d)/A(d)
Label	Building category definition
hh	High load variability & high weather sensitivity
hl	High load variability & low weather sensitivity
lh	Low load variability & high weather sensitivity
ll	Low load variability & low weather sensitivity

model, we calculate the average of the absolute value of the error terms ($|e(\mathrm{d,h})|$ or $|x(\mathrm{d,h})|$). These metrics are also applied to the average event-period values $X(\mathrm{d})$ and $E(\mathrm{d})$.

3. Building characteristics

An examination of some general characteristics of the sample of buildings is very helpful in interpreting the results of our analysis of baseline models. Here we focus on the weather sensitivity and the load variability. In this section we present the methods used to quantify these characteristics. In discussing the modeling results below, we will classify the building sample into high vs. low variability and high vs. low weather sensitivity, as shown in Table 2.

3.1. Weather sensitivity

Weather sensitivity is a measure of the degree to which building loads are driven directly by local weather. By far the most important weather variable is temperature. Weather dependence is often represented by using regression models relating hourly load to hourly temperature, possibly including lagged variables or more complex functions of temperature. The Kema 2003 report [6] investigated a number of weather regression models, but it is not clear from that study that including additional variables leads to a consistent improvement in the accuracy of the models tested. In some climates humidity may be an important factor in weather sensitivity, but for sites in California, weather behavior is likely to be dominated by dry bulb outdoor air temperature, which is the variable used here. The models tested here are based on simple linear correlation of hourly load with hourly temperature. This approach effectively rolls all other building-specific factors into the regression coefficients.

Table 2
Building sites and categorization based on weather sensitivity (ws) and load variability (var).

Site name	ROC	VAR	ws	var
Retail6	**0.97**	**0.20**	h	h
Retail4	**0.91**	**0.19**	h	h
Office2	**0.83**	**0.22**	h	h
Office3	**0.82**	**0.27**	h	h
Office/LM7	**0.77**	**0.19**	h	h
Detention facility	**0.71**	**0.24**	h	h
Office/LM1	0.65	**0.17**	l	h
Office/LM4	0.63	**0.15**	l	h
Office/LM8	0.60	**0.32**	l	h
*Museum	0.49	**0.29**	l	h
Office/Lab3	0.49	**0.18**	l	h
Office5	0.40	**0.29**	l	h
Office/LM9	0.36	**0.63**	l	h
Office/LM6	0.17	**0.96**	l	h
*School1	−0.05	**0.41**	l	h
*School2	−0.23	**0.34**	l	h
Supermarket	**0.93**	0.10	h	l
Office/LM5	**0.88**	0.11	h	l
Retail3	**0.83**	0.13	h	l
Retail5	**0.83**	0.10	h	l
Office4	**0.82**	0.14	h	l
Office/DC3	**0.79**	0.11	h	l
Office/Lab2	**0.79**	0.15	h	l
Retail2	**0.77**	0.10	h	l
Office/DC1	**0.75**	0.10	h	l
Office1	**0.75**	0.15	h	l
Office/DC2	**0.74**	0.14	h	l
Retail1	**0.71**	0.12	h	l
Office/LM2	0.64	0.11	l	l
Office/Lab1	0.61	0.13	l	l
Office/LM3	0.45	0.14	l	l
Bakery	0.01	0.11	l	l

The boldface signifies values defined as high in the categorization.

To develop an a priori sense of whether a building is likely to be weather sensitive, we use a simple and robust correlation function known as Spearman Rank Order Correlation (ROC) [11]. The ROC is obtained by first replacing each variable with its rank relative to the rest of the set and then calculating the linear correlation coefficient between the two sets of ranks. For two time series of identical length and spacing, the statistical significance of the ROC can be calculated explicitly without approximation. This metric is insensitive to the size of hourly variation in the two signals, and measures only the degree to which they tend to rise and fall together. This makes it more straightforward to compare correlation magnitudes across buildings that may have very different sized loads.

For each site, we calculate the rank order correlation between load and temperature for each hour separately for all the admissible days. We calculate the ROC coefficient in each hour separately to avoid spurious correlations driven by the daily work schedule. We also calculate an average coefficient over all the hours, which is used as an overall indicator for the building. In all cases except two the significance is greater than 95%. The two exceptions are schools which are closed from mid-June to September. The algorithm works correctly for these sites, but what it picks out is an anti-correlation between load and temperature (because load drops in the summer), and a strong random component which leads to a very low statistical significance. The cutoff for low vs. high weather sensitivity is set at a coefficient of 0.7. The ROC values by site are listed in Table 2.

3.2. Load variability

Load variability refers to how different the load profiles are from one day to another, which affects the degree to which the loads on a given day can be predicted from previous data. To quantify the variability, we use a simple measure based on the deviation of the load in each hour from an average calculated over the analysis period, in this case all the admissible days. The deviation is defined as the average value of the difference between the load in a given hour and the period average load for that hour. This is converted to a percent deviation by dividing by the period average. This variability coefficient can take on any value greater than zero, with low values indicating low variability. In order to derive a single value for each facility in our sample, we average the values calculated for each hour. Facilities are classified as either high or low variability, with the cutoff chosen at 15 percent. The variability measures for each site are shown in Table 2.

In our sample there are three buildings with non-standard schedules (indicated in Table 2 by asterisks). Two are schools that are closed during the summer as noted above. The third is a museum that is closed on Mondays and most Tuesdays. Although these schedules are perfectly predictable, they deviate from the assumption that normal operating days are Monday to Friday year-round. This results in an artificially high level of variability in load, with corresponding reduced estimate of weather sensitivity, for these sites.

4. Baseline load profile models

We tested seven baseline models for our sample of buildings, with and without the morning adjustment factor applied. These models can be loosely categorized into two groups: (1) *averaging methods*, which use some linear combination of hourly load values from previous days to predict the load on the event day (models 1 to 4), and (2) *explicit weather models*, which use a formula based on local hourly temperature to predict the load (models 5 to 7). The estimation methods tested are listed below.

For conciseness, each model is given a number followed by a short title in parentheses.

Model 1 (Average): 10-Day simple average with morning adjustment. The average of the hourly load over the 10 most recent admissible days before the event is used to predict the load on the event day. Simple average models without morning adjustment were also tested in [7].

Model 2 (Enernoc): Weighted average baseline with morning adjustment. In recent regulatory discussions on load impact estimation protocols, EnerNOC has proposed a recursive formula to predict the load on day d from predictions over a set of *N* previous days [12]. We apply this model with *N* = 20. The formula is equivalent to a weighted average of the form:

$$\mathrm{pl(d,h)} = 0.1 \times [\textstyle\sum_{(m=0,N-1)} (0.9)\mathrm{m} \times \mathrm{al(d\text{-}m,h)}] + (0.9)N \times \mathrm{al(d\text{-}}N\mathrm{,h)} \qquad (2)$$

Model 3 (3 of 10): Simple average over the highest 3 out of 10 most recent admissible days, with morning adjustment. The 3 days with the highest average load during the event period are selected from the previous 10 days, and the simple average of the load over these three days is calculated for each hour.

Model 3n (3 of 10 with no adjustment). As in Model 3 (3 of 10), the highest 3 of the most recent 10 days are averaged, but in this case no morning adjustment is applied to the estimated baseline. This is the only model presented here that does not use the morning adjustment. This baseline is model is currently used in California's Demand Bidding and Critical Peak Pricing programs and was also tested in [7].

Model 4 (5 of 10): Simple average over the highest 5 out of 10 most recent admissible days, with morning adjustment. This method is similar to the 3 of 10 method.

Model 5 (seasonal weather): Seasonal load-temperature regression model with morning adjustment. In this method, historical data is used to calculate the coefficients of a linear model of the form pl(d,h) = C1(h) + C2(h) × temperature(d,h). The coefficients are calculated using linear regression over data from all the admissible days in May to October 2006.

Model 6 (10-day weather): 10-day load-temperature regression model with morning adjustment. This method uses a linear regression model as in model 5, but the coefficients are calculated using only data from the 10 most recent admissible days prior to the event period.

A seventh model, a variation of the seasonal load-temperature regression, was also tested. In calculating the regression coefficients, instead of using all the admissible data from May to October, the data were filtered to include only those hours for which the temperature was greater than or equal to 60 F. The results obtained do not differ significantly from those for model 5 (seasonal weather), and are not included in the tables.

5. Results

5.1. Morning adjustment

Overall, we find that the morning adjustment factor substantially improves the performance of each baseline model; both in terms of reduced bias and improved accuracy [3]. A representation of the effect of applying the morning adjustment factor, using data from all models and all sites, is shown in Fig. 1. This figure is constructed as follows: For each building and each model, we calculate the probability that the relative error in the model prediction will be smaller than 5% in absolute value. This probability is estimated by counting the number of times the absolute value of *e*(d,h) is less than 5%, and dividing by the total number of values of *e*(d,h) available. This calculation is done for each model both with and without the morning adjustment. In the figure, for each building-model pair, the probability that the error will be small when no morning adjustment is applied is plotted on the horizontal axis, and the probability that the error will be small when the morning adjustment is used is shown along the vertical. For reference, the diagonal is shown on the plot as a heavy dark line. If the morning adjustment had no effect on the probability of a small error, all the points would lie along the diagonal. If a point lies above the diagonal it means that the error is more likely to be small when the morning adjustment is used. The fact that most points are above the diagonal means that in most cases the morning adjustment increases the probability that the error will be small. The linear fit shows that on average, for a given building-model pairing, the probability of small error is increased by about 25 percent when the morning adjustment is applied. Some cases are improved a great deal, others are improved only slightly, and there are a few cases where using the morning adjustment factor produces slightly larger errors.

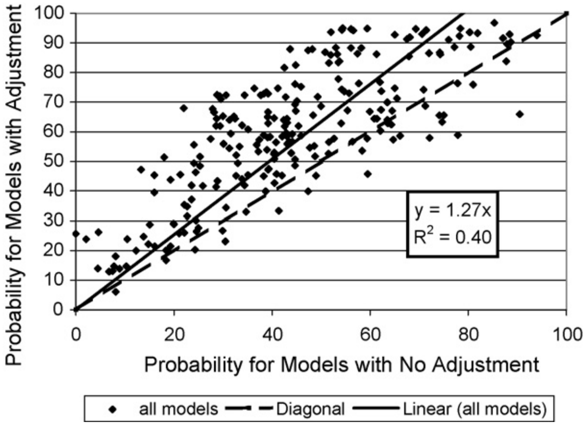

Fig. 1. Scatter plot showing how the probability that the model error will be small (less than 5% in absolute value) changes when the morning adjustment is applied. Each point corresponds to a single site–model pair. The probability that the error will be small when no morning adjustment is applied is plotted on the horizontal axis, and the probability that the error will be small when the morning adjustment is used is plotted on the vertical axis. The diagonal is shown for reference, as well as a linear fit to the data. The latter shows that applying the morning adjustment increases the probability of a small error by about 27%.

We have observed two situations where building or facility operating issues are likely to be misrepresented with morning adjustments [5]. These are related to demand response end-use strategies that begin prior to the start of the DR event (i.e., not gaming), and are important for day-ahead or other pre-notification programs. The first situation is when pre-cooling is done only on DR event days, and not on normal days. If the chiller load is higher than normal on the morning of a DR event day, the baseline load will be adjusted to a higher value than if the pre-cooling had not occurred. The adjustment reflects a demand response strategy, not the fact that the day is hotter than normal. In the second situation, we have observed industrial demand response strategies that involve reducing the end-use loads one to two hours prior to the beginning of the DR event. This is done because some industrial loads take time to "unload". In this case the morning load is lower than it would have been in the absence of a DR event, so the morning adjustment will scale the baseline down more than is appropriate. These issues suggest that some information about the building DR strategies would be very useful in assessing whether and how a morning adjustment should be applied to a baseline model.

Table 3
Bias and accuracy measures based on hourly data for each building category and baseline model. The results for a building category are defined as a simple average over the results for each building in that category. The two high schools with anomalous schedules have been excluded from the high variability, high weather-sensitivity category.

	Building category				
Variability	All	High	High	Low	Low
Weather sensitivity	All	High	Low	High	Low
Error bias measure					
Model 1 (simple average)	−1.3	−0.6	−1.7	−1.2	−0.6
Model 2 (Enernoc)	−1.4	−0.6	−1.6	−1.5	−0.7
Model 3 (3 of 10)	0.0	−0.3	0.7	−0.2	0.8
Model 3*n* (3 of 10 no adjustment)	2.2	2.3	3.7	0.4	2.0
Model 4 (5 of 10)	−0.8	−0.3	−0.6	−0.5	−0.2
Model 5 (seasonal weather)	0.4	0.8	−0.2	0.4	0.8
Model 6 (10-day weather)	0.0	0.4	0.0	−0.2	−0.5
Error magnitude measure					
Model 1 (simple average)	6.4	4.9	7.5	3.8	4.9
Model 2 (Enernoc)	6.5	4.9	7.6	3.9	4.8
Model 3 (3 of 10)	6.7	5.2	8.2	3.8	5.2
Model 3*n* (3 of 10 no adjustment)	11.3	8.1	16.2	5.8	6.4
Model 4 (5 of 10)	6.5	5.2	7.8	3.7	5.0
Model 5 (seasonal weather)	7.6	6.0	9.4	3.3	5.4
Model 6 (10-day weather)	7.1	5.4	8.9	3.7	5.7

5.2. Bias and accuracy

Table 3 summarizes the analysis results for the relative bias and accuracy among the various baseline models tested (detailed tables, including building-level results, are provided in [3]). The Table provides results for the distribution of hourly percent errors $e(\mathrm{d,h})$ between predicted and actual load by building category. The bias is measured using the median of the sample of values, and the accuracy is measured by the average of the absolute value of the error. The building category average is calculated as a simple average of the metric over all the buildings in a given category. Because the number of buildings in each category is small, these results should be taken as indicative of general trends. The building sample includes two schools which are closed during the summer months and therefore show a very high degree of variability, and poor predictability with all models under current assumptions about admissible days. These two sites are excluded from the averages for their building category.

Table 3 shows clearly that the model with no morning adjustment, model 3*n*, performs relatively poorly, and in particular it is more biased and less accurate than the same algorithm (model 3) with the morning adjustment factor included. Without the adjustment, the 3 of 10 model tends to under-estimate the event day loads, and therefore is likely to under-estimate the DR program savings. This stands in sharp contrast to previous work by Buege et al. [13], who found that the 3-of-10 model with no morning adjustment produced the highest estimates of customer baseline and the largest savings estimates for the California demand bidding and CPP tariffs. However, the load impacts from the sample of sites they evaluated were dominated by a relatively small number of large industrial customers, which presumably have low weather sensitivity, whereas our results are for commercial/institutional customers with varying degrees of weather sensitivity.

Fig. 2 provides a building-level picture of the effect of the morning adjustment for the 3 of 10 model. In this figure each point along the horizontal axis corresponds to a single building, and the bias metric is plotted on the vertical axis. The buildings are labeled and grouped according to the categorizations presented in Table 2 of high/low load variability and high/low weather sensitivity. The building category labels (hh, hl, lh, ll) are defined in Table 1. The figure shows that for most buildings if no adjustment is applied the actual load is consistently higher than the predicted load. This implies that this model will tend to predict baselines that are too low, and could therefore lead to an underestimation of the demand response impact. Applying the adjustment does not eliminate bias at the building level, but the sign is now random so that when averaged over buildings the method has a very low bias. The figure also illustrates the intuitively reasonable fact that for buildings with high load variability (categories hh and hl) the adjustment factor has a greater impact on the model results.

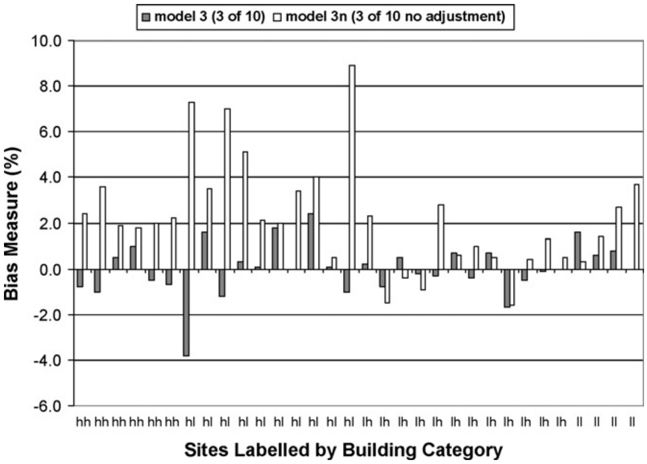

Fig. 2. Comparison of the bias measure for model 3, the average of the highest 3 of the 10 most recent days, with (grey bars) and without (white bars) the morning adjustment factor. Each point on the horizontal axis corresponds to one site. The sites are grouped and labeled by building category. The category labels are defined in Table 1.

With respect to the bias indicator, the best performance across all building categories is obtained by model 3 (3 of 10) and model 6 (10-day weather), both with morning adjustment. The seasonal weather regression model (model 5) does not perform better than the 10-day model. This may be due to the fact that the regression equation contains only linear terms, whereas the data show a slightly parabolic shape which would be better represented by a quadratic model. As noted above, in calculating averages for building categories positive and negative values of the bias metric cancel each other out. A more detailed examination of the results suggests that model 6 (10-day weather) is the only model that consistently avoids bias at the building level [3].

For the accuracy metric, the one model with no morning adjustment (model 3*n*) is the least accurate, particularly for high-variability buildings. This is intuitively reasonable, as for the proxy event days the morning load values should be a good predictor of the overall level of building activity for that day. For buildings with low variability, all models (with adjustment) perform reasonably well. For buildings with high weather sensitivity (categories hh and lh), overall the explicit weather models either improve the performance for that building or do not affect it much. This accuracy metric does not appear to discriminate as strongly among the different models as the bias metric.

5.3. Event day shed load estimates

Electric system operators and utilities with demand response programs use baseline models to estimate the customer load reduction achieved from changes to building operation during DR events. The reduction or *shed* load is a scalar quantity defined as the average over the event period of the estimated baseline load minus the measured (curtailed) event-period load. It is positive if the actual average load during the event period is lower than the baseline estimate.

Fig. 3 provides a visual representation of how the estimated load sheds vary for different models. The plot shows data for sites that participated in a series of real event days in 2005 and 2006 in California, and showed some significant demand reductions (the event days and participating sites are itemized in [4] and [5]). Sites for which no model predicts a shed of greater than 10% are excluded. In Fig. 3 each point along the horizontal axis corresponds to a single site and a single event day. As in Fig. 2, the sites are grouped according to the building category, which is indicated by the labels. The low-variability, low-weather sensitivity category (ll) is not included in the figure due to insufficient data for these event days. On the vertical axis the estimated load shed divided by the actual average load is plotted for different models. This percentage is used so that all the sites can be plotted on the same scale. For clarity, only three models are shown: model 1 (simple average, open squares), model 3 (3 of 10, triangles), and model 6 (10-day weather, filled diamonds), all with the morning adjustment. Negative values are included in Fig. 3; these correspond to cases where the model baseline values are lower than the actual load. To provide some structure to the figure, the data are sorted according to the value of the predicted shed for model 3 (3 of 10).

Fig. 3 illustrates several points. First, within a given building category a wide range of sheds are seen. This range is broader than the scatter in the different model results, and presumably represents different levels of success in achieving load reductions. Second, the relative sheds predicted by different models vary systematically with building category. The simple average model (model 1) is a smoothing operation and so one would expect it to predict lower baselines, hence lower sheds, than the other two models. This is the case for the low load variability, high weather sensitivity category (lh) and for the high load variability, low weather sensitivity category (hl), but not for sites with high load variability and high weather sensitivity (hh). For sites in the hh category, although the predicted shed can vary significantly by model, there do not appear to be any systematic differences. The weather-sensitive model (model 6) shows a tendency to predict higher sheds than the averaging models for sites in the hl category, slightly lower sheds for sites in the lh category, and no clear pattern for the hh category. Given the relatively small size of these data sets, it is not possible to say whether these tendencies are truly systematic. If they are, it is possible that for some building categories the weather-sensitive models are more appropriate, which could have a significant impact on the overall program impact evaluation and customer compensation.

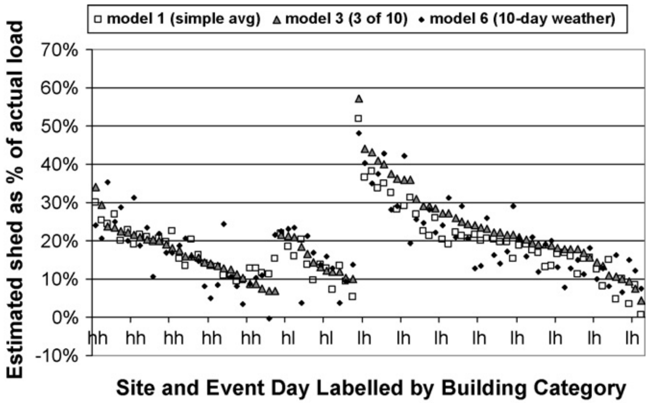

Fig. 3. Estimated load sheds for sites that participated in real event days in California in 2005 and 2006. Three models are shown: model 1 (simple average, open squares), model 3 (3 of 10, triangles), and model 6 (10-day weather, filled diamonds), all with morning adjustment. Each point along the horizontal axis corresponds to a single site and a single event day. These are sorted on the value of the shed predicted by model 3, and grouped and labeled according to the building category. On the vertical axis is the estimated load shed divided by the actual average load.

It should be noted that in this data set several of the event days occurred during multi-day heat waves [3–5]. Because event days are excluded from the set of admissible days, when events occur over consecutive days, most methods will calculate the same unadjusted baseline for each day. The adjustment factors will differ on each day, but because of changes to building operation during the event, the morning loads used to calculate the adjustment may no longer reflect the normal correlation of that building's load with that day's weather. Explicit weather models may or may not have this problem, depending on the data selection criterion. This study has not investigated multi-day events in detail, but it seems likely that model performance could differ under these circumstances.

6. Conclusions and suggestions for future work

We believe that the methods used in this study provide a statistically sound approach to evaluating the performance of different baseline load models for a building or set of buildings, provided sufficient historical data are available. The use of proxy event days expands the effective sample size and allows the methods' performance to be evaluated across a wide set of conditions for a single facility. When combined with quantitative measures of load variability and weather sensitivity, these techniques can provide a useful screening indicator to predict which types of models will perform well for a given building or facility type.

Our results show that applying a morning adjustment factor significantly reduces the bias and improves the accuracy of all the models examined in this sample of buildings. In particular, the 3 of 10 model currently used in California for several DR programs, is improved substantially here when the morning adjustment factor is applied. This finding is consistent with previous studies [7,8,13].

In our sample, models that incorporate temperature tend to improve the accuracy of the estimated baseline loads, and in cases where it does not improve the accuracy it has relatively little impact. Explicit weather models (in particular, the 10-day regression model 6) are the only model type that consistently avoids bias in the predicted loads in this sample of buildings [3].

For facilities with highly variable loads, we find that no model produces especially good results. More elaborate models do not show better performance than simple averaging methods in terms of accuracy, although bias does vary with the type of model. These types of customers can be difficult to characterize with standard approaches that rely on historic loads and weather data. It may make more sense to direct these facilities to enroll in DR programs with rules that require customers to reduce load to a firm service level or guaranteed load drop. Alternatively, additional building data could be used to incorporate some of the sources of variability directly into the model. For buildings with low load variability all the models tested perform reasonably well in accuracy.

Many demand response programs apply similar baseline estimation methods to both commercial and industrial sector facilities. The results of our study, when combined with results of other recent studies [7,8,13], suggests that DR program administrators should have flexibility and multiple options for suggesting the most appropriate baseline method for specific types of customers. Key load characteristics to be considered are weather-sensitivity (for commercial and institutional buildings but not industrial process loads) and variability of loads. For buildings with predictable but non-standard schedules, including schedule information in the selection of the admissible set would reduce the variability in the load data, and therefore improve the model performance.

Application of the methods developed here to a larger sample of buildings, covering a wider geographical area, would be very useful in determining the robustness of the results. It would also be useful to investigate whether using a more restricted proxy event set (e.g., the highest 10% of days in temperature instead of the highest 25%) leads to larger differences in model performance. The key issue is whether model performance is systematically different for real event days, which are generally not statistically typical. Larger datasets would also be needed to further investigate how to model loads during multi-day events, which are not uncommon in practice [4,5].

Acknowledgements

The work described in this report was funded by the Demand Response Research Center which is funded by the California Energy Commission, Public Interest Energy Research Program, under Work for Others Contract No. 150-99-003, Am #1 and the Office of Electricity Delivery and Energy Reliability, Permitting, Siting and Analysis of the U.S. Department of Energy under contract No. DE-AC02-05CH11231.

Researchers who contributed to this paper include Ranjit Bharvirkar and June Han at Lawrence Berkeley National Laboratory (LBNL). The authors are grateful for the support from Mike Gravely, Martha Brook, and Kristy Chew (California Energy Commission).

References

[1] FERC Staff Report, Assessment of Demand Response and Advanced Metering, Docket Number AD-06-2-000, Aug 2006.

[2] California Public Utilities Commission, Order Instituting Rulemaking Regarding Policies and Protocols for Demand Response Load Impacts Estimates, Cost-Effectiveness Methodologies, Megawatt Goals and Alignment with California Independent System Operator Market Design Protocols, OIR-07-01-041, January 2007.

[3] K. Coughlin, M.A. Piette, C.A. Goldman, S. Kilicotte, Estimating Demand Response Load Impacts: Evaluation of Baseline Load Models for Non-Residential Buildings in California, Lawrence Berkeley National Laboratory, LBNL-63728, Berkeley CA, 2008, Available at http://drrc.lbl.gov/drrc-pubsall.html.

[4] M.A. Piette, D.S. Watson, N. Motegi, S. Kilicotte, P. Xu, Automated Critical Peak Pricing Field Tests: Program Description and Results, Lawrence Berkeley National Laboratory, LBNL-59351, Berkeley, CA, 2006, Available at http://drrc.lbl.gov/drrc-pubsall.html.

[5] M.A. Piette, D.S. Watson, N. Motegi, S. Kilicotte, Automated Critical Peak Pricing Field Tests: 2006 Pilot Program Description and Results, Lawrence Berkeley National Laboratory, LBNL-62218, Berkeley, CA, 2007, Available at http://drrc.lbl.gov/drrc-pubsall.html.

[6] M.L. Goldberg, G. Kennedy Agnew, Protocol Development for Demand-Response Calculations: Findings and Recommendations, KEMA-Xenergy, CEC 400-02-017F, 2003.

[7] Working Group 2 Demand Response Program Evaluation—Program Year 2004 Final Report, Quantum Consulting Inc. and Summit Blue Consulting, LLC, 2004.

[8] Evaluation of 2005 Statewide Large Nonresidential Day-ahead and Reliability Demand Response Programs, Quantum Consulting Inc. and Summit Blue Consulting, LLC, 2006.

[9] National Oceanic and Atmospheric Administration, National Climatic Data Center, http://www.ncdc.noaa.gov/oa/ncdc.html.

[10] California Department of Water Resources, California Irrigation Management Information System, http://www.cimis.water.ca.gov/cimis/welcome.jsp.

[11] William H. Press, Brian P. Flannery, Saul A. Teukolsky, T. William, Vetterling. Numerical Recipes in FORTRAN 77, Cambridge University Press, 1992.

[12] D. Kozikowski, A. Breidenbaugh, M. Potter, The Demand Response Baseline, v.1.75, EnerNOC OPS Publication, 2006.

[13] A. Buege, M. Rufo, M. Ozog, D. Violette, S. McNicoll, Prepare for Impact: Measuring Large C/I Customer Response to DR Programs, ACEEE Summer Study on Energy Efficiency in Buildings, Monterey, CA, August 2006.

[14] N. Motegi, M.A. Piette, D.S. Watson, S. Kilicotte, P. Xu, Introduction to Commercial Building Control Strategies and Techniques for Demand Response, Lawrence Berkeley National Laboratory, LBNL-59975, Berkeley, CA, 2007, Available at http://drrc.lbl.gov/drrc-pubsall.html.