Building Technologies & Urban Systems Division
Energy Technologies Area
Lawrence Berkeley National Laboratory

Benchmarking Demand Flexibility in Commercial Buildings and Flattening the Duck – Addressing Baseline and Commissioning Challenges

Jingjing Liu[1], Lili Yu[1], Rongxin Yin[1], Mary Ann Piette[1], Marco Pritoni[1], Armando Casillas[1], Monica Neukomm[2], Amir Roth[2]

Lawrence Berkeley National Laboratory
US Department of Energy

Energy Technologies Area
August 2022

https://doi.org/10.20357/B7M89Q

This work was supported by the Assistant Secretary for Energy Efficiency and Renewable Energy,
Building Technologies Office, of the US Department of Energy
under Contract No. DE-AC02-05CH11231.

Benchmarking Demand Flexibility in Commercial Buildings and Flattening the Duck – Addressing Baseline and Commissioning Challenges

Jingjing Liu, Lili Yu, Rongxin Yin, Mary Ann Piette, Marco Pritoni, Armando Casillas,
Lawrence Berkeley National Laboratory
Monica Neukomm, Amir Roth, U.S. Department of Energy

ABSTRACT

With the transition from our traditional electric grid to a cleaner grid with renewable power generation, there is a need to enable building loads to be flexible. Load shedding and shifting will be essential for flattening the "Duck" for decarbonization. This paper explored the trend in the timing of DR events as a reflection of the grid's needs using recent four years of event data from 203 retail stores in 11 states. The events are becoming significantly shorter with 2-hour duration being the most popular; shifting to late afternoon and early evening is another trend beyond California.

Benchmarking will be essential for accounting DF as a reliable grid resource. This paper addresses a challenging aspect of benchmarking – inaccuracies in counterfactual baseline methods can introduce significant DF metrics variations in addition to weather and building characteristics related factors. The conventional "10/10" with adjustment baseline method has inherent limitation by design for load shifting applications. Therefore, it is imperative to identify alternative methods. This study compared three hourly regression baseline methods with "10/10" methods using two groups of commercial buildings that participated in DR programs: (1) 121 big-box retail stores, and (2) 11 office buildings in CA. The 14-day hourly outdoor temperature regression method was found to produce least error in the tested datasets and is promising for load shifting.

The paper also pointed out that commissioning issues can also be a significant barrier for achieving consistent DF performance, which building managers and utilities should be aware of.

Introduction

Building Demand Flexibility (DF) Benchmarking

The U.S. Department of Energy (DOE) launched a major research initiative in 2018 named "Grid-interactive Efficient Buildings" (GEB) to better understand the role of building "demand flexibility" (DF). GEB defines DF as "the capability of DERs to adjust a building's load profile across different timescales" and has been focused on "Load Shed"[1] (or "*Shed"*) and and "Load Shift"[2] (or "*Shift*") (Neukomm et al. 2019). Building "demand flexibility", also

[1] Load Shed: the ability to reduce electricity use for a short time period and typically on short notice. Shedding is typically dispatched during peak demand periods and during emergencies (Neukomm et al. 2019).
[2] Load Shift: the ability to change the timing of electricity use. In some situations, a shift may lead to changing the amount of electricity that is consumed.

known as "load flexibility" or "energy flexibility" has been increasingly recognized as an important grid resource in recent several years along with other distributed energy resources (DERs) (Hledic et al. 2019, Satchwell et al., 2021). Today's grid is changing – when DF resources will be called upon by the grid operators and the duration of these called grid events are both changing. This paper uses a significant dataset to investigate this trend.

Despite the potential, there is a lot to be understood about DF performance in order for the grid operators to treat it as a reliable grid resource. On the other hand, building managers and operators also want to understand how their buildings' DF compare with other buildings and what influences DF performance so as to harvest the full value of participating in DF programs. To fill the gap of standardized DF metrics, several research studies have proposed metrics that measure the size and value of DF resources at the building level or system level (Liu et al. 2020, Miller and Carbonnier 2020, Langevin et al. 2021, Aduda et al. 2018, Cai and Braun 2019). One building level metric particularly useful for DF benchmarking was "demand decrease intensity [W/ft^2]"[3], or "*shed intensity*" (Liu et al. 2020) because the floor area normalized metric allows comparison among buildings.

The chosen baseline method can have significant impact on metrics values and, therefore, on DF benchmarking results. To support robust benchmarking results, this paper explores the accuracies of multiple baseline methods.

Baseline Challenge in Measuring Real Buildings' DF

Creating reliable counterfactual baselines for each *shed* or *shift* event day can be challenging. Today's DR programs typically adopt day-matching or weather-matching type of baseline methods for commercial building customers (Bode and Ciccone 2017). Some researchers have also explored using more sophisticated baseline methods for DR purposes. A recent study (Granderson et al. 2021) tested using three variations of a sophisticated piecewise linear regression baseline method, called "time of week and temperature (TOWT)" (Mathieu et al. 2011a; Granderson et al. 2016), along with traditional day-matching and weather-matching baseline methods to predict hourly electricity consumption during peak load periods in commercial buildings. The study used 12 months of hourly data for 453 electricity meters to predict the 10 highest load days of the year for each meter; those meters did not participate in DR. It found that all tested baseline methods significantly under-predict consumption by 4.5-18.7% during those peak load periods on 1,104 prediction days. Although the TOWT method was proven accurate for predicting annual energy use, it did not test to be more accurate than the conventional baseline methods with the dataset in the referenced paper. It highlighted the need for DR baseline methodology research.

Furthermore, there are additional challenges to creating accurate baseline for load shifting as compared to load shedding (i.e. traditional DR). Applying a "morning adjustment factor" to a traditional "10/10" baseline[4] for **shed** events can often improve the baseline accuracy (Piette et al., 2006; Coughlin et al., 2009, Goldberg et al., 2013). It does so by comparing the average load

[3] Defined as the average kW demand reduction during a shed event or window normalized by the building floor area. The area normalization allows comparison of shed metrics across buildings of different sizes or types.

[4] "10/10" baseline: a type of day-matching baseline method which takes the average demand during DR event hours over the previous 10 eligible baseline days (excluding weekends, holidays, DR event days, and none-operation days)

of the couple of hours before a shed event on the event day vs. that of the 10 previous non-event days. However, when a building **shift** loads, for example, by pre-cooling the space, the "morning adjustment factor" can introduce bias beyond its intended function to the baseline load shape and potentially over-predict "shed" (i.e. load decrease) and under-predict "take" (i.e. load decrease). Therefore, there is research needed for identifying or creating a baseline method that is suitable for calculating DF metrics for load shifting. In search for a suitable baseline method for load shifting, researchers should be aware of the limitation of exclusively using non-DR event days to test the accuracies of different baseline methods because such bias introduced by using adjustment factors during "load take" periods would not be detected using non-event day load profiles.

In this paper, we test variations of an existing hourly linear regression baseline method, which have been used in multiple previous DR field studies with positive results (Mathieu et al., 2011b, Kim et al., 2013) but have not been applied to a larger commercial building dataset for accuracy test.

What Influences Demand Flexibility?

A real building's DF performance is complex. Here we take *shed* as an example to discuss influential factors, but it's similar for *shift*. The "benchmarking demand decrease intensity" metric during a *shed* event is subject to the influence of at least the following factors which are categorized as:

(1) **Building characteristics:** e.g. building type, vintage, operating hours, and climate zone;
(2) **DF metric attributes:** e.g. DF strategy, DF event duration, time-of-day, day-of-week, year, baseline method, weather condition;
(3) **Stochastic factors:** e.g. stochastic phenomenon in occupancy, internal loads, and ***business operations***, ***commissioning issues*** in equipment and controls.

Among the above three categories of factors, the influence from building characteristics is static in the sense that the factors do not change day-to-day or week-to-week. Ideally, buildings with similar characteristics will be grouped together for benchmarking. The factors in the second category are variable from one event to another and will influence the metric value for the same building. In other words, each DF metric value should specify a set of *Attributes* ("DF Metrics Attributes") that are associated with it; they together define a building's DF performance. The data type of factors in the first two categories are often either categorical or discrete numerical values. The third category of factors are stochastic in nature and often influence DF performance in the background.

As discussed above, the first two categories of factors are covered in benchmarking by grouping building characteristics and attaching attributes to metric values. This paper does not discuss them in detail. However, it is worth noting that in using these characteristics or attributes related factors to analyze building DF metrics such as investigating correlations, it is important to understand that sometimes the stochastic factors can have significant impact on the results especially when the dataset is relatively small. In this paper, we discuss some examples of such factors related to business operations and commissioning issues in a later section.

Objectives

There are three objectives in this study:
(1) Identify trends in the timing of DR events to support understanding the changing needs of the grid and timing of DF resources needed;
(2) Investigate an alternative baseline method for load shedding and load shifting;
(3) Identify examples of commissioning issues that can influence DF performance besides building characteristics and events related factors.

Case Studies

To accomplish the three objectives identified above, we use two datasets of real buildings that participated in utilities DR programs. They are included for different purposes. The main dataset is 203 big-box retail stores located across 11 states in the US, mainly concentrated in the follow ASHRAE climate zones[5]: 2A (Hot-Humid), 3A (Warm-Humid), 3B (Warm-Dry), 4A (Mild-Humid), 5A (Cold-Humid). Among the 203 sites, 121 of them were selected for the baseline analysis due to incomplete data issues in the other sites. These stores have participated in various DR programs through their local utilities and the recent four years of data was available. Therefore, this dataset was used to explore the trend of DR events called by the utilities as an indicator of the changing needs of the grid. In addition, this dataset is also used to investigate the accuracies of multiple baseline methods for *shed* because these stores deployed load shedding strategies.

The second dataset is 11 medium office buildings in California from a previous field study (Yin et al. 2008 a, Yin et al. 2008 b). These office buildings have deployed pre-cooling strategy before the DR events which is considered *shift* strategy. We used this dataset to understand how different baseline methods work for office buildings, which have different daily load shape and weekly patterns from retail stores. These baseline methods are then used to estimate counterfactual load shapes on event days to provide insights to how each method works for *shift*.

More details about each building group is provided below.

Retail Stores

These 121 big-box retail stores typically range from about 90,000 – 190,000 sqft in size and were mostly constructed before 2012 with various renovation statuses. They have similar envelopes (concrete wall and small window areas) and HVAC system type (packaged rooftop units [RTU] with variable speed fans).

Whole-building power data was collected at 15-min intervals for 2018-2021 (4 years total), which included two years before and two years since the COVID-19 pandemic. While the pandemic had a temporary impact on shopping traffic, it did not impact the HVAC operations significantly according to the building staff.

The stores generally deployed the following two DR strategies at each site. First, about half of the lights were turned off in the sales areas (the actual percentage of lights may vary by site.) Second, a number of the RTUs deployed a global zone temperature adjustment (GTA)

[5] See ANSI/ASHRAE Standard 169-2013, Climatic Data for Building Design Standards

strategy[6]. It is worth noting that the stores had upgraded its control algorithm for the sites between the DR seasons in 2019 and 2020, which would account for a significant part of the DR performance change before and after 2020. This impact is considered more significant than the pandemic by the building staff. After the control upgrade, the sites would consistently increase space temperature set point up by ~3 ºF during a DR event, whereas the temperature adjustment varied significantly by each RTU based on the previous control algorithm. In addition, the selection of RTUs that participate in DR events at each site have also changed when the control upgrade was implemented. Considering these complexities and lack of site-by-site level of information, it is not in the scope of this paper to interpret the root causes of the metrics results presented.

Medium Office Buildings

The basic building characteristics of these 11 medium office buildings in California have been provided in a previous study (Liu et al. 2020) along with details of the DR strategy used. In short, 2-3 ºF of GTA control was used during DR events (all events were 12-6pm in 2008) and the majority of the sites also implemented 2 ºF pre-cooling before the event to shift load.

In this paper, the 15-min whole-building power raw data was used to reconstruct different counterfactual load shapes using multiple baseline methods to compare the baseline model accuracies using non- DR event days and visually compare them in time-series plots for DR event days.

The Changing Trends of DR Events

We aggregated the DR event date and time window information during 2018-2021 for the 203 retail stores to explore the trend in the timing of DR events. We examined three aspects of the event timing including event duration, start hour, and month of the year, and aggregated them by year and by state, respectively (see **Figure 1** and **Figure 2**). It is worth noting that in this dataset, most of the sites and states included are located in the Middle Atlantic, South Atlantic, East North Central, West South Central, and the Pacific regions of the US.

Figure 1 shows that there is a clear increasing trend of 2-hour long events. The events are generally starting between 2-4pm and it appears that the number of events starting at 4pm is increasing although there was a modest scale back in 2021. Most events are concentrated in the summer season during June to August.

The average number of events called at these retail stores in each state is not evenly distributed. **Figure 2** shows that DR events were more frequently called in the following states: OH, VA, IL, PA, and TX. It also shows that some states such as CA have a relatively even distribution of event duration compared to other states dominated by 2-hour events. Events in CA are generally called later in the evening during 4-6pm, which is consistent with the "duck curve" phenomenon.

[6] Raising zone temperature setpoints "globally" across a number of zones in the building during a *Shed* event. It is common for buildings to couple it with "pre-cooling" by lowering the zone temperature setpoint for some time before the *Shed* in order to achieve deeper demand shed and provide better comfort during the event.

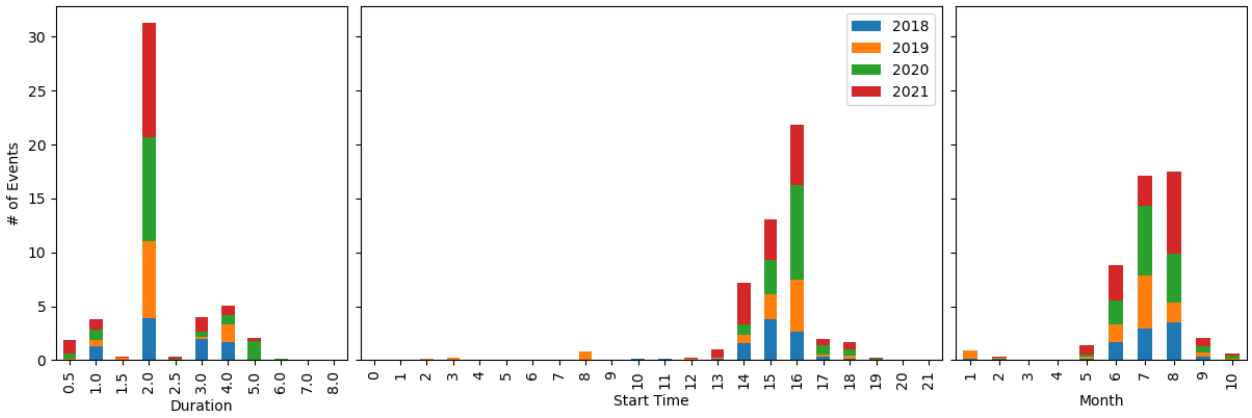

Figure 1: Average Number of DR Events Called per Site in 2018-2021 for a Group of 203 Retail Stores – Trend by Year (Left: Event Duration; Middle: Event Start Time; Right: Month) (Note: the number of events have been normalized by the number of sites in each state)

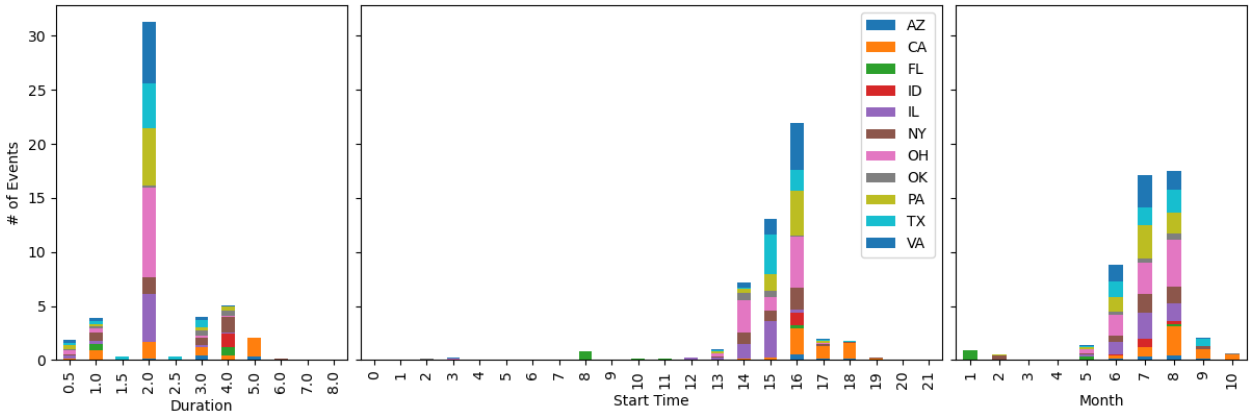

Figure 2: Average Number of DR Events Called per Site in 2018-2021 for a Group of 203 Retail Stores – Trend by Year (Left: Event Duration; Middle: Event Start Time; Right: Month) (Note: the number of events have been normalized by the number of sites in each state)

Baseline Method Comparison and Discussion

When calculating DF metrics, significant variations can be introduced to the metrics results using different baseline methods. Such variations directly affect how one would interpret the buildings' DF performance which may or may not match the reality. Therefore, using a robust baseline method to calculate DF metrics is critical to a relatively accurate understanding of a buildings' true performance and taking appropriate consequential actions based on such understanding. We describe the methodology used in comparing five baseline methods and discuss the results below.

Baseline Methods Descriptions

In this paper, we use the two datasets described above to compare a total of five baseline methods in two categories as shown in **Table 1** below.

Table 1. Summary of Five Baseline Methods for Comparison

Short Name	Baseline Method Category	Method Variation	Description
"10/10"	day-matching	original	(see footnote 1)
"10/10 adj."	day-matching	adjustment factor	First 3 of the 5 hours before an event is used for adjustment. Any adjustment is limited to ±20%
"OAT reg. (14 days)"	Hourly linear regression, single-variable	short period	14 days of non-holiday days before the event; predict using outdoor temperature alone
"OAT reg. (30 days)"	Hourly linear regression, single-variable	medium period	30 days of non-holiday days before the event; predict using outdoor temperature alone
"OAT + DoW reg."	Hourly linear regression, multi-variable	medium period	4 weeks of non-holiday days before the event; predict using outdoor temperature and day of week

In order to compare their accuracy, we use all five baseline methods to predict the hourly energy consumption for each of the 121 sites on all of the non-event days. We then calculate two error metrics, *Normalized Mean Bias Error* (or "nMBE") and *Coefficient of Variance of the Root Mean Square Error,* or "CV(RMSE)", against the actual measured hourly consumption. The equations for these two error metrics are provided below (**Eq. 1** and **Eq. 2**). In general, the smaller their absolute values the more accurate the prediction model is. In nMBE definition, positive errors imply the baseline method under-predicts the consumption and negative errors imply over-predicting. ASHRAE Guideline 14 (ASHRAE 2014) recommends that in calibrating whole-building energy simulation, the CV(RMSE) should be no greater than 30% and the nMBE should be within ±10%. These thresholds are high; therefore, we focus on comparing their relative values among the multiple baseline methods rather than the threshold values.

$$CV(RMSE) = \frac{\sqrt{\frac{1}{n-p}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}}{\bar{y}} \times 100 \qquad \text{(Eq. 1)}$$

$$nMBE = \frac{\frac{1}{n-p}\sum_{i=1}^{n}(y_i-\hat{y}_i)}{\bar{y}} \times 100 \qquad \text{(Eq. 2)}$$

where yi is the actual metered value, $\hat{y}i$ is the predicted value, $\bar{y}$ is the average of the yi, and n is the total number of data points, with p = 1.

Considering previous studies (Granderson et al. 2021) found various baseline methods are prone to under-predicting hourly consumption during peak load periods, we filtered periods in the non-event days which have conditions similar to events so that we are able to examine the errors during these event-like periods vs. other periods. For the retail stores dataset, we used 12-9 PM and outdoor air temperature (OAT) greater than 80ºF as the filtering criteria. This approach may be useful for future studies.

Retail Buildings (Load Shed)

We applied the above five baseline methods to the 121 retail stores described earlier. **Table 2** shows the average CV(RMSE) and nMBE calculated from the predicted vs. actual hourly consumption during filtered periods (12-9pm, OAT>80ºF) and the remainder periods on all non-DR event days in 2018-2021. These results are also shown in **Figure 3.**

It shows that all of the five methods produced errors much below the thresholds set by ASHRAE Guideline 14. More importantly, by separating all hours on non-event days into "filtered" and "remainder" periods, the nMBE results reveal that the "10/10" and "10/10 adj." methods under-predict during the filtered periods (by ~5% and 1% respectively) and over-predicts during the remainder periods. The three regression baseline methods all slightly under-predict the hourly consumption during both filtered and remainder periods by ~1% or less, and the "OAT reg. (14 days)" method had the smallest positive nMBE. The "10/10 adj." and "OAT reg. (14 days)" methods had the smallest CV(RMSE). Therefore, the "OAT reg. (14 days)" was considered the best baseline method for this dataset overall. Adding day-of-week to the regression model did not improve accuracy.

This results indicate that temperature is an important driver for cooling load in these retail stores, which explains the low errors for the OAT regression methods. However, there are other factors that drive the total building load to be higher or lower for a temporary period of time, which is why shorter model training period (14 vs. 30 days) has a small advantage in accuracy. Such factors are likely store operations related although this cannot be confirmed with limited information at hand. The operations on weekdays vs. weekends were not significantly different.

Table 2. Error Metrics of Different Baseline Methods Applied to a Group of 121 Retail Stores, Calculated for Filtered vs. Remainder Periods on Non- DR Event Days in 2018-2021

Model	nMBE		CV(RMSE)	
	Filtered	Remainder	Filtered	Remainder
10/10	4.7%	-2.1%	7.4%	11.5%
10/10 with adjustment	1.2%	-1.2%	5.5%	8.9%
OAT Regression (14 days)	0.5%	0.1%	6.4%	9.8%
OAT Regression (30 days)	0.7%	0.4%	6.7%	10.1%
OAT + DoW Regression	0.8%	1.5%	7.6%	12.3%

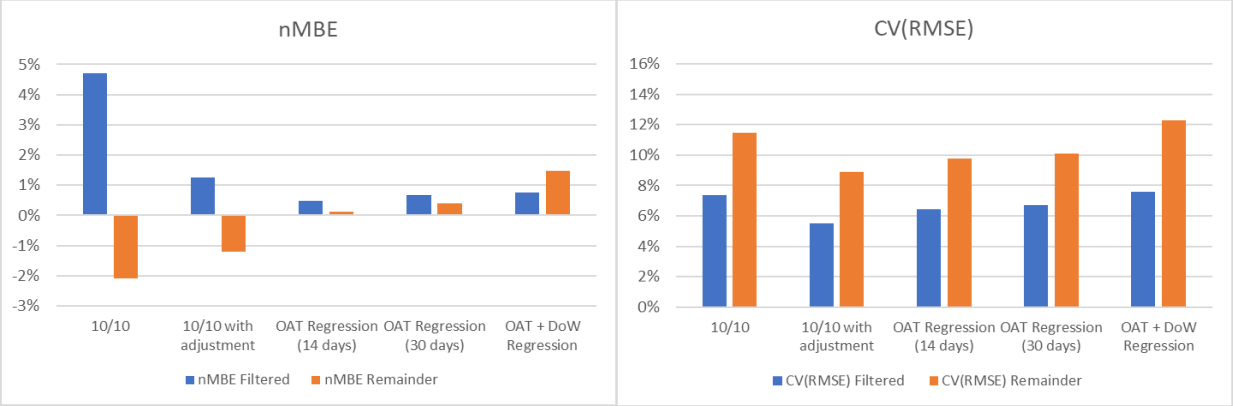

Figure 3: Error Metrics (Left: nMBE; Right: CV[RMSE]) of Different Baseline Methods Applied to a Group of 121 Retail Stores, Calculated for Filtered vs. Remainder Periods on Non- DR Event Days in 2018-2021

It is important to note that it may not be reliable to make conclusions of baseline method comparisons solely based on the above error metrics because of the averaging effect across many days and hours. We suggest researchers review time-series plots of a significant sample of DR

event and non-event days to make sure the daily load profiles are generally consistent to what the error metrics suggest before drawing conclusions. For this building group, we selected a couple of 24-hour time-series plots as examples to compare the predicted counterfactual load profiles from various baseline methods with the actual whole-building power, as shown in **Figure 4**. **Figure 4-a** shows an example of actual vs. predicted baseline load profiles on a DR event day where the "10/10" method predicts the lowest and the "10/10 adj." also predicts significantly lower than the three regression-based methods. **Figure 4-b** shows an example of a non-event day where the three regression-based methods capture the afternoon load drop due to weather much more closely than the "10/10" and "10/10 adj." methods.

These examples along with the error metrics results above show that the hourly regression based baseline methods are promising for DR applications in the big-box retail building type tested. We further test them using a separate dataset for office buildings in the next section.

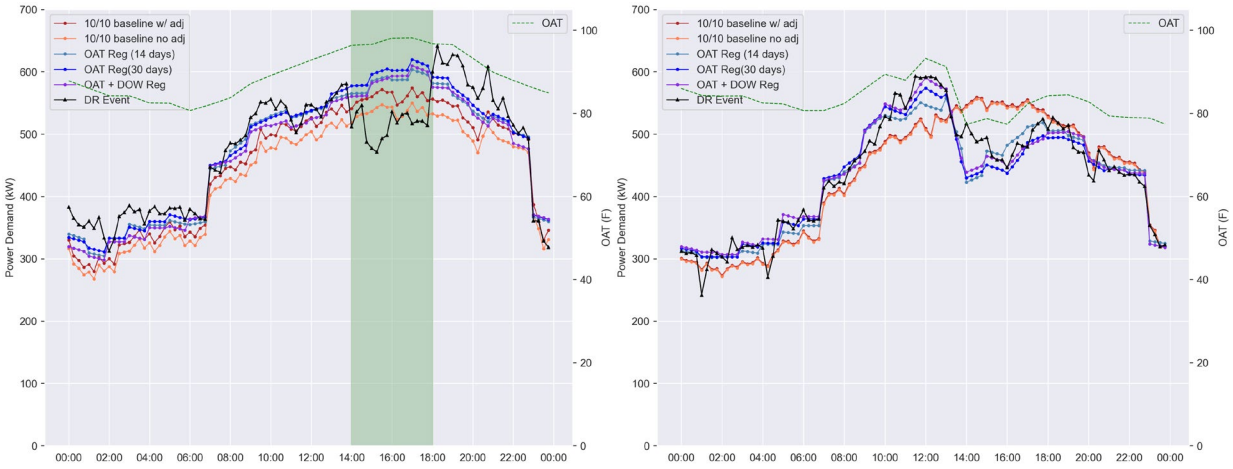

Figure 4: Time-Series Plot of Actual Power and Different Counterfactual Baseline Load Shapes on a DR Event Day (Left: 4-a) and a Non-Event Day (Right: 4-b) for a Selected Retail Store

Figure 5 shows an example of the "demand decrease intensity [W/ft^2]" metric for 2-hour shed events for the 121 retail stores using the 14-day hourly OAT linear regression model. It shows that the group of buildings achieved 0-0.75 W/ft^2 through lighting and HVAC zone temperature adjustment controls strategies. As discussed in an earlier section "What Influences Demand Flexibility", many factors including the baseline method can contribute to such significant variation in shed metrics among these buildings despite similar control strategies. The more accurately we can improve the baseline method, the more meaningful correlation analysis and conclusions can be done to understand the influence of other factors.

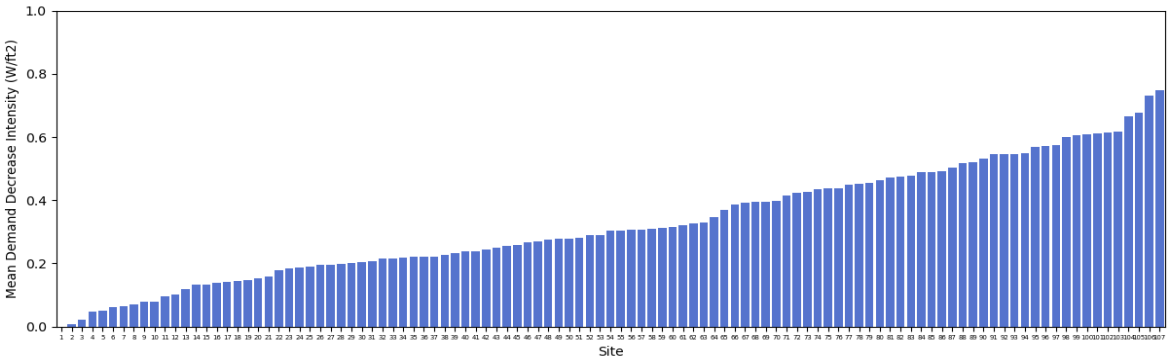

Figure 5: Demand Decrease Intensity [W/ft^2] Metrics of a Group of 121 Retail Stores, Calculated for 2-Hour DR Events in 2018-2021 Using 14-day Hourly OAT Regression Baseline Method

Medium Office Buildings (Load Shift)

We applied the same five baseline methods to the 11 medium office buildings introduced earlier. **Table 3** and **Figure 6** shows the average CV(RMSE) and nMBE calculated from the predicted vs. actual hourly consumption during filtered periods (12-6pm, OAT>80ºF) and the remainder periods on all non-DR event days during June through September in 2008.

Again, all of the five methods produced errors much below the thresholds set by ASHRAE Guideline 14 (30% for CV(RMSE) and ±10% for nMBE). The nMBE results reveal that all methods under-predict the hourly consumption during the filtered and remainder periods (by ~1-2%) except that the "10/10 adj." method slightly over-predicts (by 0.5%) during filtered periods. The "10/10 adj." method produced the smallest nMBE in absolute terms followed by the "10/10" and "OAT reg. (14 days)" method. These three methods are also measured with the smaller CV(RMSE) in the same order (10-14%).

Besides testing these methods with a different building type, one important point to note with this building group is that if we solely use non-event days to measure the accuracy of these baseline methods, we would have concluded that "10/10 adj." was the best because it produced the smallest errors measured by nMBE and CV(RMSE). However, because this group of buildings had implemented pre-cooling and shifted some load from afternoon to morning, the "10/10 adj." method will introduce a bias by design, which would not be detected on the non-event days. Therefore, the "OAT reg. (14 days)" and "10/10" methods with similar error results were considered better baseline methods for this dataset overall. Adding day-of-week to the regression model did not improve accuracy for this dataset as well.

Table 3. Error Metrics of Different Baseline Methods Applied to a Group of 11 Medium Office Buildings, Calculated for Filtered vs. Remainder Periods on Non- DR Event Days in 2008

Model	nMBE		CV(RMSE)	
	Filtered	Remainder	Filtered	Remainder
10/10	1.4%	2.1%	12.6%	13.4%
10/10 with adjustment	-0.5%	1.0%	10.5%	8.9%
OAT Regression (14 days)	1.5%	1.4%	13.6%	11.5%
OAT Regression (30 days)	1.9%	1.6%	13.7%	11.6%
OAT + DoW Regression	2.0%	1.7%	15.0%	12.7%

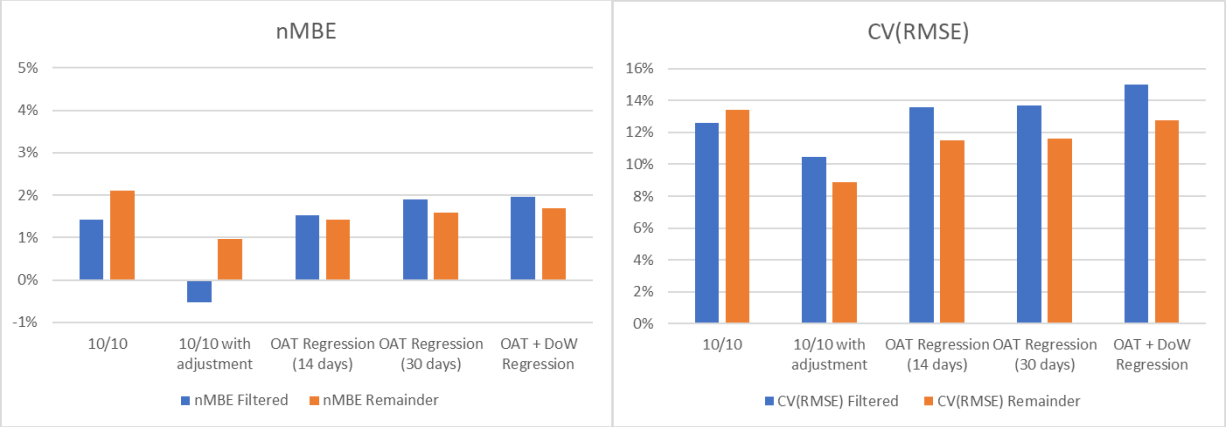

Figure 6: Error Metrics (Left: nMBE; Right: CV[RMSE]) of Different Baseline Methods Applied to a Group of 11 Medium Office Buildings, Calculated for Filtered vs. Remainder Periods on Non- DR Event Days in 2008

As noted earlier, it is important to review time-series plots before drawing conclusions on baseline accuracies. For this building group, we selected a time-series plot of a DR event day and a non-event day below in **Figure 7**. **Figure 7-a** shows that on a DR event day, the "10/10 adj." method predicts significantly higher than the other baseline methods due to the higher cooling load from the morning pre-cooling strategy. As mentioned earlier, this would not be revealed by using only non-event days to evaluate the accuracy of baseline methods. **Figure 7-b** shows that the "10/10 adj." method also over-predicted on a non-event day for a different reason. In this example, the building experienced higher cooling load on Monday mornings in preparation for occupancy after the weekend shutdown (or reset); this effect can also significantly bias the "10/10 adj." baseline prediction if a DR event happens on Mondays. Both examples suggest that "10/10 adj." should not be used for load shifting. **Figure 7-b** also shows the "OAT + DoW reg." method captured morning "warm up" effect the best although this period is generally outside the DR events time window.

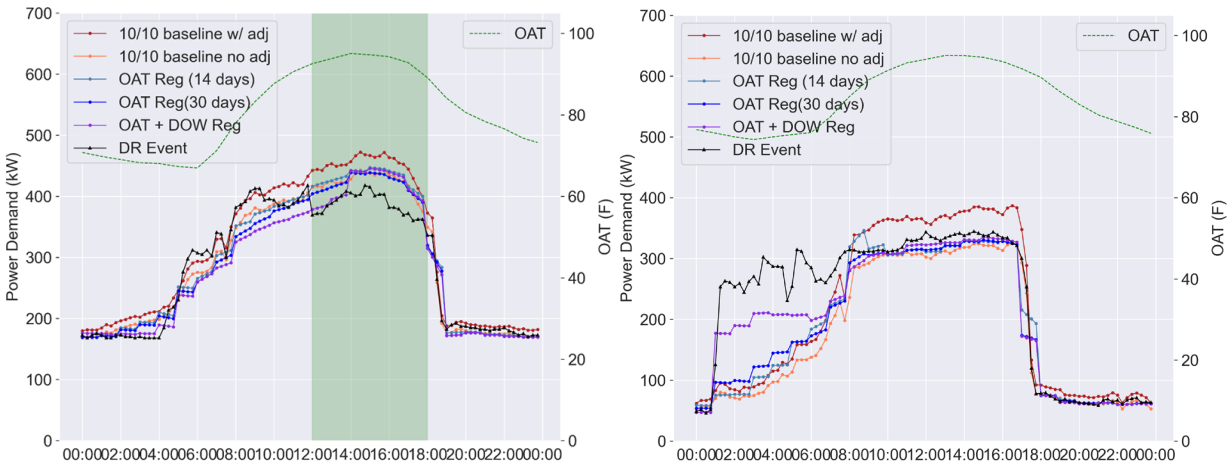

Figure 7: Representative Time-Series Plots of Actual Power and Different Counterfactual Baseline Load Shapes on a DR Event Day (Left: 7-a) and a Non-Event Day (Right: 7-b) for a Group of Medium Office Buildings

Summary of Baseline Method Comparison

In this paper, we tested five baseline methods in two categories (see **Table 1**) using two groups of buildings of different types (big-box retail stores and medium office buildings), both of which have participated in DR events. We primarily used two error metrics from ASHRAE Guideline 14 calculated for the non- DR event days to compare the accuracies of these methods, but also reviewed a significant sample of time-series plots of both event and non-event days to ensure that they are consistent with the error metrics results overall. In calculating error metrics, those periods on non-event days with warmer weather and higher frequency of DR event calls were separated from the other hours in order to better understand the predicting errors during these "event-like" periods.

The retail store dataset containing 121 sites with 4 years of data provided significant support for using a single-variable (OAT) hourly linear regression model as an alternative to the conventional "10/10" baseline method with or without adjustment factors. The 14-day variation of the hourly OAT linear regression model performed slightly better than the 30-day variation or adding a day-of-week variable in the two datasets tested. Further experiments with additional building types and datasets would be beneficial for fully understanding the performance of these regression based methods. Despite a smaller dataset, the office building group reveals that using error metrics on non-event days alone to compare baseline methods has blindspots - the adjustment factor in "10/10" baseline method has inherent limitations in some load shifting applications (shifting load to shortly before a DR event.)

Operations and Commissioning Issues Influence DF Metrics

Some types of commercial building tend to follow weekday/weekend operations and have relatively consistent load shape patterns week-to-week. Office buildings are such an example because they are driven by regular occupancy. In other building types, such as retail stores, "events" in business operations (e.g. promotional events, loading of refrigerated products, maintenance of refrigerated cases, etc.) can cause a significant increase in electricity load profile.

Such load increase will have an impact on DF metrics if a DR event is called on these days. They can also increase the baseline load profile if included in the baseline data sample for a following DR event day. These can create challenges for accurately calculating and interpreting DF metrics results because operational "events" are not always regular or known to people who perform the metrics analysis.

Besides business operations related factors, equipment and control issues in the energy-consuming systems (e.g. HVAC, lighting) can also have significant impact on DF metrics. In this paper, we refer to these issues as "commissioning" for simplicity. Below is not a comprehensive list but rather some examples related to the GTA and lighting DR strategy used in the building groups in this paper.

- Delayed equipment repair: e.g. when a RTU fails and is not repaired timely, the DF metric will be impacted;
- Equipment retrofit: e.g. when lighting fixtures are replaced with more efficient ones, the DF from lighting load will likely decrease;
- Sensor or control issues: e.g. if a RTU is short-cycling, then the load shed from implementing GTA will likely be compromised;
- Adjacent zones: e.g. when multiple RTUs serve a shared zone or zones without physical barriers, if one unit implemented GTA then the other unit(s) are likely to experience increased load so the overall shed result may be less than anticipated;
- Sub-metering: e.g. if sub-metered building loads such as HVAC are used to calculate DF metrics instead of whole building power, the data source needs to be updated when equipment is replaced or status in DR participation changes. Gaps in such communications can lead to misleading metrics results.

Operations and commissioning issues related variations in baseline load profiles and DF metrics are complex to identify and address, and therefore need more research.

Summary and Next Steps

In this paper we address three aspects related to DF benchmarking: (1) exploring the trend of DR events timing as a reflection of the grid's potential needs for DF resources; (2) investigating the accuracies of multiple variations of hourly linear regression models in anticipation for non-conventional baseline method needs for load shifting; (3) raising the issue that business operations and building systems commissioning can significantly influence DF and deserves attention. The key findings, limitations and future work needed in each of these three areas are summarized below.

(a) DR Event Trend. Recent four years of DR event data from 203 retail stores in 11 states show that the events are becoming shorter and late afternoon 2-hour events were most popular. This exploration is limited to the utility and third party DR programs that these retail stores were enrolled in, and therefore, does not reflect the larger trend of DR events timing in the US. Larger and more devise datasets can make such trend study more meaningful.

(b) Baseline Methods for DF. In testing three variations of hourly linear regression model based baseline methods using two significant real building datasets, we found that they

overall introduce smaller errors during "event-like" periods compared to the conventional "10/10" baselines. Among the three variations, the regression model that uses only OAT as an independent variable and 14 qualified days from prior to the prediction day yielded the least error although its advantage over the 30-day and 2-variable (OAT and day-of-week) variations were small. This means such regression baseline methods could be promising for load shifting applications where the conventional "10/10" with adjustment factor baseline will most likely introduce systematic biases. This test was done on two relatively homogeneous datasets - one larger for retail stores across several climate zones and one much smaller for office buildings in CA. Larger and more diverse datasets should be used for additional testing of the candidate baseline method before drawing conclusions. The filtering method used for examining errors was a contribution of this study.

(c) Operations and Commissioning Issues. We categorize the many influential factors of DF metrics into building characteristics, DF metric attributes, and stochastic factors. Among the stochastic factors, business operations and commissioning issues are often overlooked and yet can significantly impact DF metrics. They need to be well understood and addressed in order for DF to serve as a reliable grid resource. We provided several examples to raise awareness. Future DF metrics studies should be connected with synergistic fault detection and diagnostics studies to explore how to smoothly deliver building DF reliably to the grid.

Acknowledgements

This work was funded by the Assistant Secretary for Energy Efficiency and Renewable Energy, Building Technologies Office, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

References

ASHRAE. 2014. ASHRAE Guideline 14-2014 – Measurement of Energy, Demand and Water Savings, American Society of Heating, Refrigeration and Air Conditioning Engineers, Atlanta, GA.

Bode, J. and Ciccone, A. 2017. California ISO Baseline Accuracy Assessment (November 2017). CAISO Baseline Accuracy Working Group. (Retrieved on 3/23/2022 from https://www.caiso.com/Documents/CaliforniaISOBaselineAccuracyAssessmentNexant.pdf)

Coughlin, K., Piette, M.A., Goldman, C., and Kiliccote, K. (2009). "Statistical Analysis of Baseline Load Models for Non-Residential Buildings,". Energy and Buildings, 41(4), 374-381. https://doi.org/10.1016/j.enbuild.2011.08.020

Neukomm, M., Nubbe, V., and Fares, R. (2019). "Grid-interactive Efficient Buildings Overview (December 2019)".

Goldberg, M.L., and Agnew, G.K., (2013), Measurement and Verification for Demand Response, Prepared for the National Forum on the National Action Plan on Demand Response: Measurement and Verification Working Group, FERC.

Granderson, J., Sharma, M., Crowe, E., Jump, D., Fernandes, S., Touzani, S., and Johnson, D. (2021). "Assessment of Model-Based Peak Electric Consumption Prediction for Commercial

Buildings". Energy and Buildings, Volume 245, 2021, 111031, ISSN 0378-7788, https://doi.org/10.1016/j.enbuild.2021.111031.
Granderson J., Touzani S., Custodio J., Sohn M., Jump D. ,Fernandes S., 2016. "Accuracy of Automated Measurement and Verification (M&V) Techniques for Energy Savings in Commercial Buildings." Applied Energy, 173, pp.296-308.
Hledik, R., Faruqui, A., Lee, T., and Higham, J. (2019). The National Potential for Load Flexibility Value and Market Potential Through 2030. The Brattle Group.
Jihyun Kim, J., Yin, R., and Kiliccote, S. (2013). Automated Price and Demand Response Demonstration for Large Customers in New York City using OpenADR. In 2013 International Conference for Enhanced Building Operations. Montreal, Quebec, October 8-10, 2013
Langevin, J., Chioke, B., Satre-meloy, A., Speake, A., Present, E., Wilson, E. J. H., and Satchwell, A. J. (2021). Article US building energy efficiency and flexibility as an electric grid resource and flexibility as an electric grid resource. Joule, 5(8), 2102–2128. https://doi.org/10.1016/j.joule.2021.06.002
Li, H., Wang, Z., Hong, T., and Piette, M. A. (2021). Energy flexibility of residential buildings : A systematic review of characterization and quantification methods and applications. Advances in Applied Energy, 3(June), 100054. https://doi.org/10.1016/j.adapen.2021.100054
Liu, J., Yin, R., Pritoni, M., Piette, M. A., and Neukomm, M. (2020). Developing and Evaluating Metrics for Demand Flexibility in Buildings : Comparing Simulations and Field Data. 267–280.
Mathieu, J. L., Callaway, D. S., & Kiliccote, S. (2011a). Variability in Automated Responses of Commercial Buildings and Industrial Facilities to Dynamic Electricity Prices. Energy and Buildings, 43(12), 3322–3330. https://doi.org/10.1016/j.enbuild.2011.08.020
Mathieu, J. L., Price P., Kiliccote S., Piette M.A.. (2011b). "Quantifying changes in building electricity use, with application to Demand Response." IEEE Transactions on Smart Grid 2:507- 518.
Miller, A. and Carbonnier, K. (2020). New Metrics for Evaluating Building-Grid Integration. 2020 ACEEE Summer Study on Energy Efficiency in Buildings (Virtual Conference).
Piette, M. A., Watson, D., Motegi, N., Kiliccote, S., & Xu, P. (2006). Automated Critical Peak Pricing field tests: Program description and results. Lawrence Berkeley National Laboratory. LBNL-62218.
Satchwell, A., Ann Piette, M., Khandekar, A., Granderson, J., Mims Frick, N., Hledik, R., Faruqui, A., Lam, L., Ross, S., Cohen, J., Wang, K., Urigwe, D., Delurey, D., Neukomm, M., & Nemtzow, D. (2021). A National Roadmap for Grid-Interactive Efficient Buildings. U.S. DOE Building Technologies Office.
Yin, R., Xu, P., Piette, M. A., & Kiliccote, S. (2010). Study on Auto-DR and pre-cooling of commercial buildings with thermal mass in California. Energy and Buildings, 42(7), 967–975. https://doi.org/10.1016/j.enbuild.2010.01.008
Yin, R., Xu, P., and Kiliccote, S. (2008). Auto-DR and Pre-cooling of Buildings at Tri-City Corporate Center. 2008. Lawrence Berkeley National Laboratory. LBNL-3348E.