Seeing the (Short-Term) Future: Assessment of Demand Response M&V Baseline Methods

Eliot Crowe, Jessica Granderson, Samuel Fernandes,
Lawrence Berkeley National Lab, Berkeley, CA
Mrinalini Sharma, David Jump, Devan Johnson,
kW Engineering, Oakland, CA

ABSTRACT

Demand response (DR) programs are becoming increasingly important for utilities looking to avoid system load constraints. In anticipation of extreme grid demand, DR programs offer incentives for customers to reduce load within a time window on a specific day (a DR "event").

The most common methods for quantifying building-level load reduction during DR events rely on simple averaging algorithms using hourly load and temperature data from the days preceding the DR event to create a counterfactual baseline. Regression-based methods are now being applied to hourly meter data through advanced measurement & verification to quantify annual savings for energy efficiency projects, but are not typically employed for DR. A question arises as to how regression-based methods compare with established averaging methods for quantifying DR load reductions; specifically, how accurately can regression-based methods develop a counterfactual baseline prediction against which load reduction can be quantified? To address this question, an objective comparison/test method must be developed, as no standard baseline accuracy test exists.

This paper summarizes the development and application of a method to compare DR baseline prediction accuracy of regression-based and averaging algorithms. We tested variants of three baseline modeling approaches using a data set of thousands of hypothetical DR event days using real buildings' data from multiple regions in the United States. Our results indicate that all methods evaluated would tend to understate a counterfactual baseline, thereby understating achieved load reductions from DR programs, and that the regression methods tested do not offer a notable advantage over traditional averaging methods.

Introduction

In recent decades, demand response (DR) programs have evolved around two basic approaches: rate-based and incentive-based (Chai et al. 2019). California is an example of a state where the role of DR has grown considerably over the past two decades. In 2003 the California Energy Commission designated DR as being first in the "loading order" (the order in which resources are to be deployed), along with energy efficiency. As a result, the California Public Utilities Commission (CPUC) set a goal to meet 5% of the electric system's annual peak energy demand with DR by 2007 (whereas previously DR had only been occasionally used and considered as a kind of "insurance policy") (Jarred 2014). As of 2017, almost 19 million utility customers were enrolled in DR programs across the United States (FERC 2019).

Under incentive-based DR approaches, utility customers can receive significant financial incentives to reduce electric load during times of peak grid stress (typically referred to as a DR "event"). For example, the Eversource ConnectedSolutions DR program offers $35 per average kilowatt reduction for DR events that are called during summer months, with an expectation of no more than eight events in that time period (Eversource 2020).

It is important to quantify the impacts of incentive-based DR programs, both at the individual building level (for calculating incentive payments) and at the aggregate level (for programs or regions), to ensure that incentives are correctly calculated and to better understand the consistency of the delivered load reductions.

Background

The foundation for quantifying temporary load changes at the individual building level is to gather electricity consumption data prior to the DR event (the baseline period) and use it to create "counterfactual" load predictions, i.e., estimates of what the load would have been during the event period in the absence of the DR strategies deployed. Program evaluations performed at the aggregate level have a broader selection of established methods available for quantifying impacts (including the use of comparison groups), but may employ building-level counterfactual load predictions in some circumstances.

Literature Review

Much prior work has been conducted to assess methods for predicting building-level counterfactual energy consumption for commercial buildings; examples of this prior work are summarized below.

A California-based 2017 study (Bode and Ciccone 2017) assessed 36 permutations of three different DR baseline calculation methods, applied to large aggregations of building loads (as opposed to the loads for individual buildings):

- Control groups, where a group of meters with statistically similar electricity consumption during the baseline period are used to determine the counterfactual consumption during the event period for a group of residential DR customers.
- Weather-matching, where non-event baseline days with similar ambient temperature conditions are selected for each meter and data are averaged.
- Day-matching, where a subset of non-event days in close proximity to the event day are identified and their load data are averaged to produce baselines for an individual meter.

Additional multiplicative adjustments were made to the weather-matching and day-matching algorithms, based on the difference between predicted and actual load during pre-event or post-event hours. Baseline prediction accuracy was quantified using two metrics for assessing prediction bias and precision: mean percent error (MPE) and the coefficient of variation of the root mean squared error (CV[RMSE]). The study recommended calculation parameters for each of the three approaches tested, asserting that, for the California program dataset tested, multiple baseline rules can deliver sufficiently unbiased and precise baselines for pooled aggregates of buildings, including weather-matched and day-matched algorithms.

A study commissioned by the PJM[1] Load Management Task Force assessed several DR baseline approaches (including averaging and regression approaches), analyzing a total of 36 baseline calculation methods (KEMA 2011). The methods assessed included several types of adjustment for day-of-event conditions, including load additive adjustment, load ratio adjustment, weather sensitive adjustment, and no adjustment. The PJM results show that predictive accuracy can vary based on weather-responsiveness of load and the timing/season of the event window, and that adjustment of load estimates based on day-of-event conditions is highly beneficial. The study recommended four methods (all with additive adjustment) where median bias value across all meters analyzed was at or close to zero:

[1] PJM is a regional transmission organization (RTO) that coordinates the movement of wholesale electricity in all or parts of 13 states and the District of Columbia (www.pjm.com).

1. Prior-day baseline and current day meter data.
2. Day-matched with a prior 10 days' average.
3. High four days of most recent 45 days.
4. Day-matched with middle four of prior six days.

A 2013 study examined a number of DR baseline estimation methods used by utilities and electrical system operators across the United States and evaluated them in terms of accuracy and bias levels. They acknowledged the possibility of both bias and random error, and described four main strategies for addressing those issues: (1) perform baseline method assessment studies, (2) make operational adjustments (e.g., de-rate DR savings to avoid overcounting), (3) make adjustments to program rules, and (4) treat the DR program as an iterative process, adapting the program measurement and verification (M&V) approach based on ongoing results and the customer mix (Goldberg and Agnew 2013).

A 2002 study for the California Energy Commission tested DR baseline prediction accuracy for a variety of calculation methods (including averaging and regression) and found that additive adjustments were generally required to compensate for underestimation of load during hypothetical events. The study noted several potential challenges with this type of adjustment if applied to real DR events, such as the possibility of building owners gaming results by deliberately increasing building loads prior to the DR event. Further, the study noted that the baseline estimation applied to any given building needs to be tailored to unique circumstances such as the weather-sensitivity of its load, and whether the event is occurring in summer or winter months (Xenergy 2002).

Similar to Xenergy, Grimm 2008 also noted a need for DR baseline methods to minimize the risk of gaming (e.g., with a short notice period prior to a DR event, a customer could not deliberately inflate their consumption prior to the event; Grimm 2008), and found that methods using multiple pre-event days reduced the risk of gaming.

Advanced M&V methods have emerged over the past decade, employing hourly or sub-hourly data and sophisticated modeling approaches to quantify energy efficiency annual savings with a high degree of accuracy (Franconi et al. 2017). The "time of week and temperature" (TOWT) model is a piecewise linear regression that has been well documented in the literature (Mathieu et al. 2011; Granderson et al. 2016). The TOWT model and its variants also have been incorporated into utility program efficiency M&V and industry tools as an accepted method (CalTRACK 2018; Granderson et al. 2019; Crowe et al. 2019). Some of the first uses of this model targeted DR applications (Kiliccote et al. 2010; Mathieu at al. 2011; Price et al. 2015). Price et al. (2015) assessed the predictive accuracy of a more complex variation of the TOWT model, with a custom adjustment based on model residuals for recent non-prediction days. A cross-validation test of the studied model, assessing peak day predictions from 12:00 pm to 6:00 pm, showed median bias of less than 4% ("baseline percent error," where a positive value indicates the predictions were higher than actual consumption), compared to 6% for a day-matched 10-day algorithm and 5% for the TOWT model.

Academic literature on methods for predicting commercial buildings' energy consumption are common, but rarely focus on predictive accuracy for timescales aligning with DR and using whole building electricity consumption data.

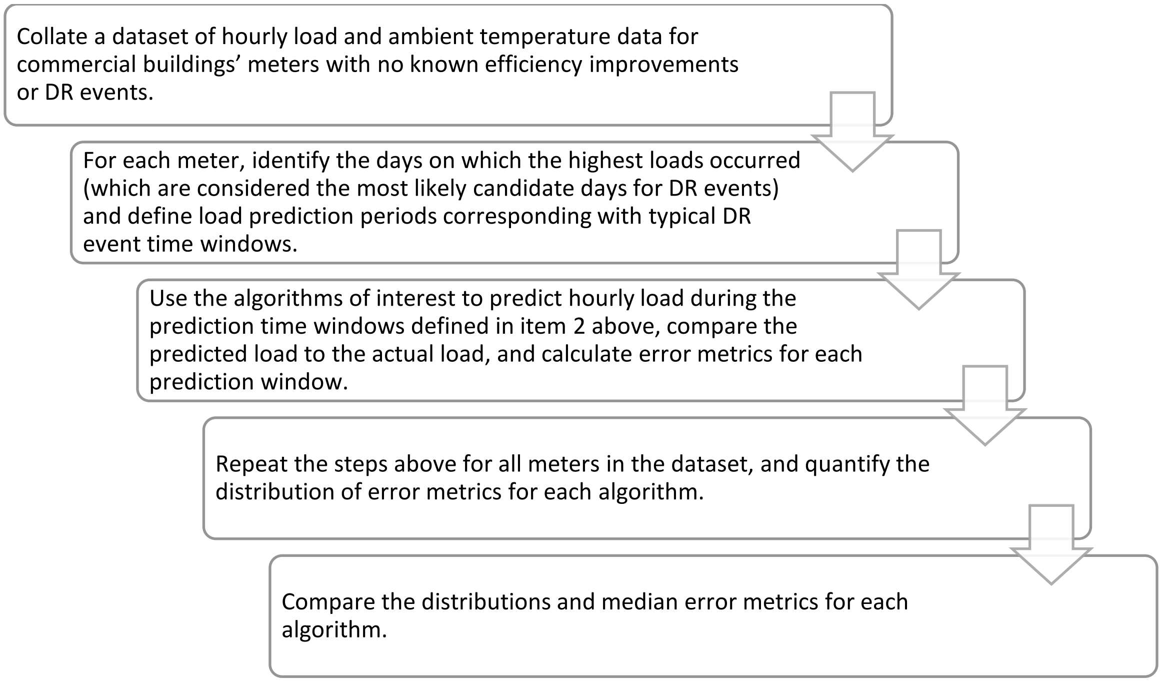
Study Overview

This study complements the body of prior work by evaluating whether a regression model that has proven accurate for predicting annual energy use is also accurate in predicting short-duration peak loads, when compared to methods that are commonly used in today's DR programs. It presents predictive accuracy results using interval meter data drawn from several regions of the United States, for eight analysis algorithms and three different time periods for over a thousand peak prediction days.

The specific research questions answered in this work were: (1) How does the advanced M&V regression-based approach compare to the established averaging methods? (2) Does the duration and timing of the DR event window have a significant impact on the prediction accuracy? and (3) Are there notable differences in the distribution of prediction accuracy results across a large population of meters when employing different baseline prediction methods?

Method

The DR baseline predictive accuracy assessment presented in this paper is based on a five-step process, as shown in Figure 1.



Collate a dataset of hourly load and ambient temperature data for commercial buildings' meters with no known efficiency improvements or DR events.

For each meter, identify the days on which the highest loads occurred (which are considered the most likely candidate days for DR events) and define load prediction periods corresponding with typical DR event time windows.

Use the algorithms of interest to predict hourly load during the prediction time windows defined in item 2 above, compare the predicted load to the actual load, and calculate error metrics for each prediction window.

Repeat the steps above for all meters in the dataset, and quantify the distribution of error metrics for each algorithm.

Compare the distributions and median error metrics for each algorithm.

Figure 1. DR baseline predictive accuracy assessment process

This study was targeted at commercial buildings. Three prediction windows were tested under this study: 10:00 am to 6:00 pm, 12:00 pm to 6:00 pm, and 1:00 pm to 4:00 pm. These were selected to allow for comparison and their selection acknowledges that DR events may occur during different time windows depending on region, generation mix, and weather conditions.

Figure 2 illustrates the baseline energy consumption data (orange) used by one prediction algorithm for one test case (i.e., a specific time window on one prediction day); this example is for a TOWT model using seven weekdays prior to the prediction day as baseline data. Figure 1 also shows the associated prediction window (10:00 am to 6:00 pm) on the event day (July 18) and plots the actual consumption (red) and the predicted values (green) from the TOWT model based on local ambient temperature data for each hour.

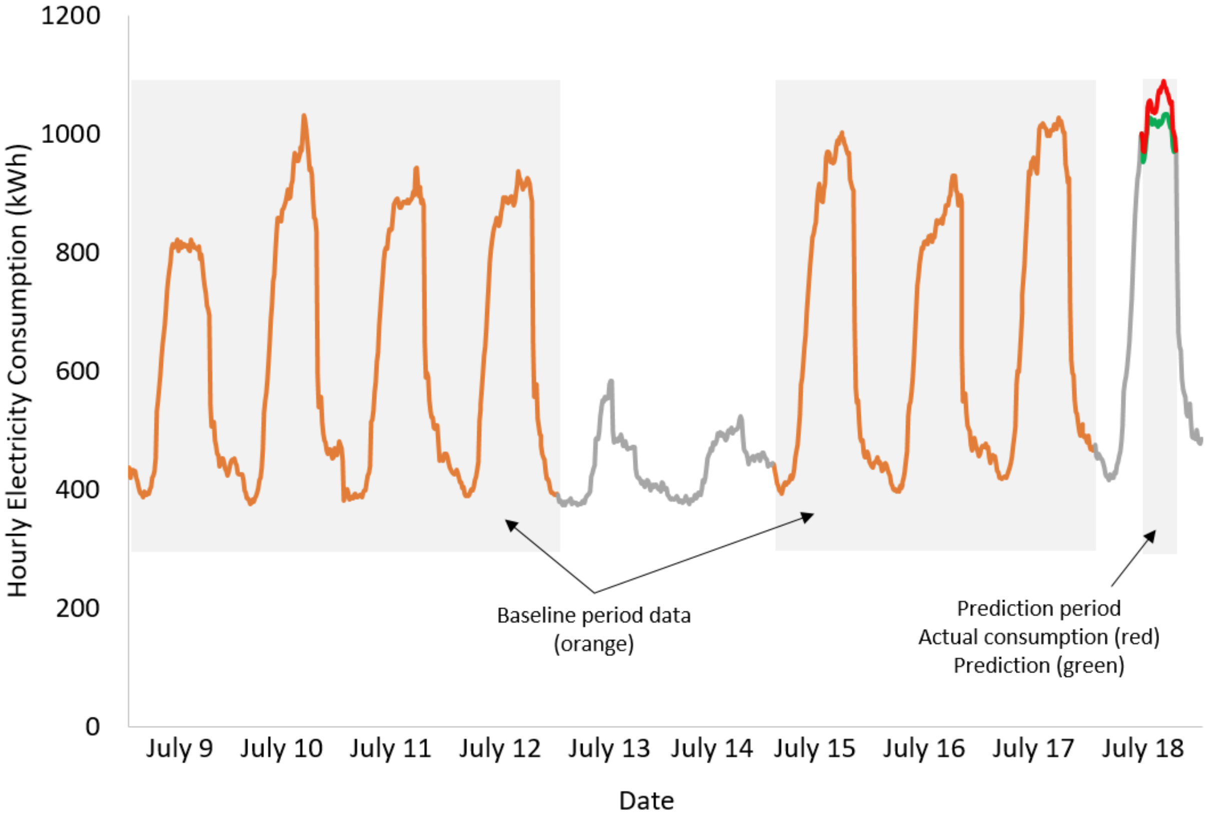


Figure 2. Example plot showing the prediction period (10:00 am to 6:00 pm on July 18) and baseline data used for one prediction algorithm studied.

Data Preparation

The test dataset covered 12 months of hourly electric consumption (kilowatt-hours, kWh) and hourly outside air-dry bulb temperature. The test data were selected from an existing dataset available to the researchers, drawn from 453 commercial buildings where no known energy efficiency projects or DR events had occurred. The data covered three U.S. Building America climate zones (Baechler et al. 2015): Marine, Cold, and Mixed-Humid. The test data were intentionally diverse in terms of region, consumption, and property type, to allow for assessment of algorithms across a diverse set of conditions. All data were cleaned of obvious erroneous values, such as unrealistically high or low temperature.

After removing holidays and weekends for each meter, the ten days with the highest maximum daily load were identified and selected as test cases ("prediction days") for this study. Any candidate day that did not have a sufficient history of data to satisfy *all* of the baseline methods was excluded; for example, the weather-matching algorithm required 90 days' worth of data prior to the prediction day. The result was 1,104 prediction days that were used in the analysis, a sufficiently large quantify to determine overall performance and variability of the algorithms' predictive capabilities.

Peak Prediction Algorithms

Three peak prediction algorithms were assessed in this study, including two averaging algorithms and a regression method. Each of these methods is described below.

Averaging Methods. Two averaging methods were selected based on the best results reported in Bode and Ciccone 2017:

- *Day-Matching:* Baseline data drawn from the 10 working days immediately prior to the event day.
- *Weather-Matching:* Baseline data drawn from the four days out of the 90 days prior to the event with maximum temperature closest to the maximum temperature of the event day.

For both day-matching and weather-matching algorithms, for each hour of the event day, the corresponding hours from the baseline data are averaged to calculate hourly predictions for the event window. These two algorithms were selected for this study as two contrasting options that had been shown to perform well and are in current use.

Regression Method. TOWT was selected for this study as an industry-accepted regression model. TOWT is a piecewise linear model where the predicted energy consumption is a combination of two terms that relate the energy consumption to the time of the week and the piecewise-continuous effect of the temperature. In previous studies (e.g., Granderson et al. 2016) the TOWT model was shown to be accurate at predicting annual consumption using hourly data, equaling or outperforming other M&V industry standard models. The TOWT model uses time of the week and outside air temperature as independent variables, and can be configured to add weighting to data toward the end of the baseline period (i.e., closer to the peak prediction window being studied). The TOWT model variants tested under this study were:

1. 7 baseline days, no weighting.
2. 70 baseline days, 14 days weighted.
3. 70 baseline days, 10 days weighted.

Adjustments for Day-of-Event Conditions. As noted earlier, adjustment methods have been developed to account for weather impacts to load on peak days; specifically, that peak days are likely to see higher temperatures than the baseline days preceding them. These methods are based on observed load before and/or after the event window. The adjustment approach documented in Bode and Ciccone 2017 was used in this study. Adjustments were calculated by comparing actual and predicted loads during hours prior to the prediction window ("adjustment hours") and using that information to scale the predictions during the prediction window (see Equation 1). Adjustment hours were selected with a buffer period of two hours from the prediction window (e.g., for the prediction window 12:00 pm–6:00 pm, adjustments were based on loads between 8:00 am and 10:00 am).

$$Adjustment\ Ratio = \frac{Actual\ total\ kWh\ during\ adjustment\ hours}{Algorithm's\ predicted\ kWh\ during\ adjustment\ hours} \quad (1)$$

Adjustment ratio caps applied in this work followed the recommendations in Bode and Ciccone 2017: 40% for weather-matching, and 20% for day-matching. A 40% adjustment cap was also applied for one TOWT variant. Table 1 lists each of the algorithms and variants tested.

Table 1. Peak Prediction Algorithms Tested

Algorithm	Variant*	Abbreviation

Day-Matching	Unadjusted	DMU
	Adjusted	DMPA
Weather-Matching	Unadjusted	WMU
	Adjusted	WMPA
Time-of-Week-and-Temperature (TOWT)	7-day baseline (no weighting)	UWTOWTU(7.0)
	7-day baseline (no weighting) (adjusted)	UWTOWTPA(7.0)
	70-day baseline (14-day weighting)	UWTOWTU(70.14)
	70-day baseline (10-day weighting)	UWTOWTU(70.10)

* Adjustments were applied to all algorithms except for the weighted TOWT models, which were excluded due to timing and resource constraints.

Assessment Metrics

Normalized mean bias error (NMBE) and the coefficient of variation of the root mean squared error (CV[RMSE]) have been used in prior work to assess accuracy of advanced M&V models (Granderson et al. 2016). NMBE and CV(RMSE) are also familiar to M&V practitioners as model fitness metrics, featuring in resources such as ASHRAE Guideline 14 (ASHRAE 2014). Equations 2 and 3 below define NBME and CV(RMSE) respectively, where y_i is the actual metered load value, $\hat{y}_i$ is the predicted load value, $\bar{y}$ is the average of the y_i, and *N* is the total number of data points.

$$NMBE = \frac{\frac{1}{N}\Sigma_i^N(y_i - \hat{y}_i)}{\bar{y}} \times 100 \tag{2}$$

$$CV(RMSE) = \frac{\sqrt{\frac{1}{N}\Sigma_i^N(y_i - \hat{y}_i)^2}}{\bar{y}} \times 100 \tag{3}$$

For this study metrics were calculated based on model predictions of data not used in the model creation, a process known as *cross-validation* or *out-of-sample testing.* NMBE and CV(RMSE) values closer to zero indicate more accurate predictions. Bias (NMBE) may be positive or negative, with positive values indicating underprediction (i.e., predicted values lower than actual values). When applying NMBE to assess model fitness, model specification is the primary source of error/bias, and an NMBE value of zero is achievable. When applying NMBE for out-of-sample testing, the data not used in model creation introduces additional potential for error. For example, the out-of-sample data may be taken from a time period after a building occupancy change. This study was designed to mitigate this potential error in two ways: through selection of a large dataset, to counteract the possibility of including some buildings with operational changes; and by focusing mainly on median NMBE results, to eliminate skew effects from outliers. It should be noted that this study is intended to provide an objective comparison between different algorithms, as opposed to expecting zero bias for any individual algorithm.

While CV(RMSE) and NMBE values were calculated for all test cases, only NMBE results are reported here, since the broad conclusions are the same for both metrics and NMBE results are more easily interpreted.

Results

Figure 3 shows a single prediction window for each tested algorithm. The plot shows actual hourly meter readings and the algorithms' predictions for the hours between 10:00am and 6:00pm on a single prediction day. This provides a visual example as context for the results that follow (this is illustrative, not a typical or average result), which summarize predictive accuracy across all the models for each of the 1,104 prediction days.

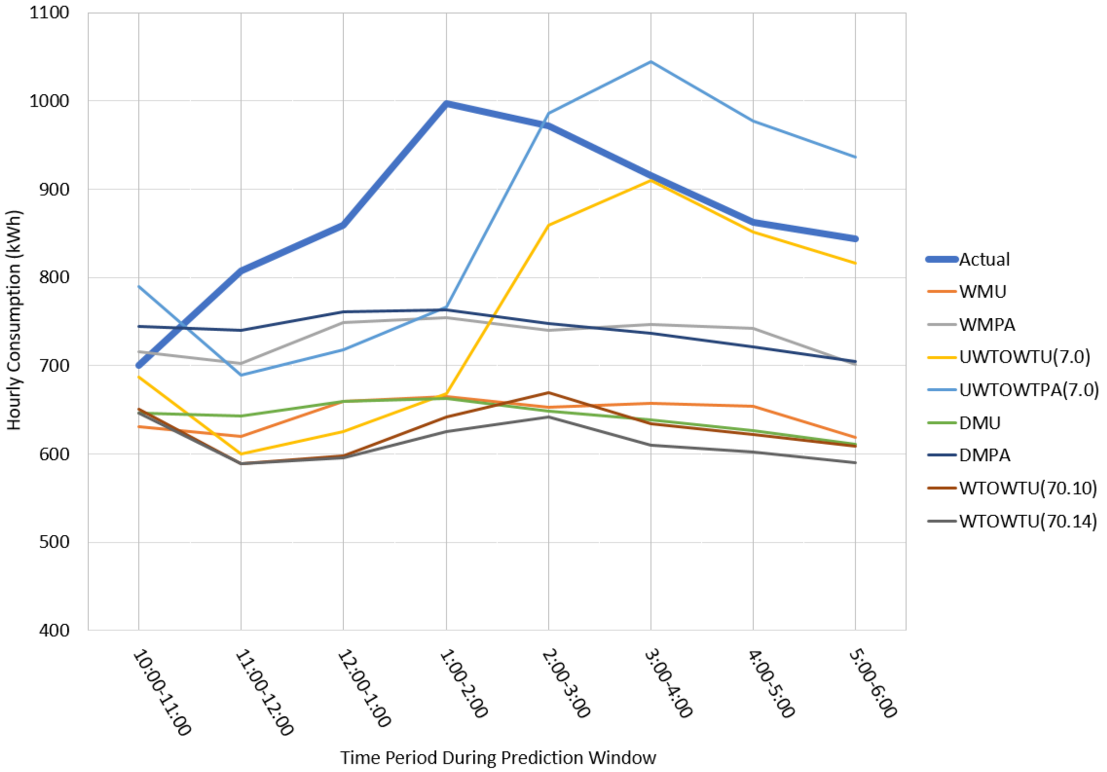


Figure 3. Example prediction window (10:00 am to 6:00 pm) for a single meter, illustrating the predictions from each of the tested algorithms compared with actual consumption.

Figure 4 shows the distribution of NMBE results for each prediction algorithm, and for the three prediction windows: 10:00 am to 6:00 pm, 12:00 pm to 6:00 pm, and 1:00 pm to 4:00 pm. The box and whisker plots indicate the 10th, 25th, 50th, 75th, and 90th percentile values.

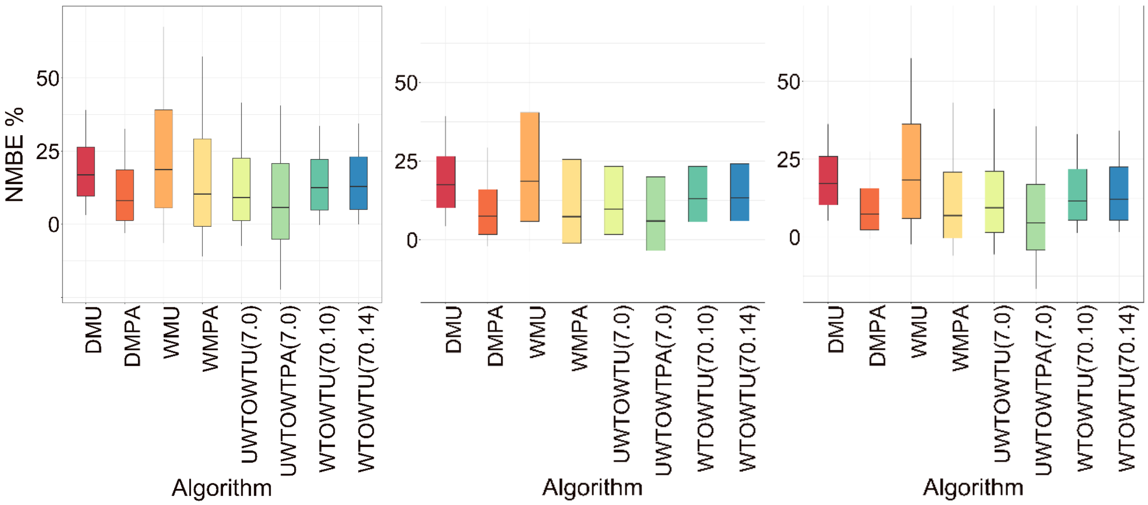


Figure 4. NMBE results distribution for 10:00 am – 6:00 pm (left), 12:00 pm – 6:00 pm (center), and 1:00 pm – 4:00 pm prediction windows

The results indicate no significant difference in predictive accuracy for the three prediction windows, and no consistent pattern in terms of which prediction window saw the highest or lowest median values. Figure 4 also shows significant overlap in distribution between all the algorithms.

The results also show wide distribution for all methods (with WMU algorithm having the widest distribution), with significant overlap between algorithms. All median NMBE values were biased in a positive direction (see Table 2), indicating underestimation of the load during peak hours. The lowest median bias (4.5%) was observed for the unweighted, adjusted TOWT regression with the 1:00 pm–4:00 pm prediction window; this median value is still considered high relative to published literature. Additionally, the results show that the application of adjustments had a significant effect on NMBE. Adjustments were applied in three cases (DMPA, WMPA, and UWTOWTPA), across the three event time windows; in six out of nine of these cases the adjustment reduced the median NMBE value by over 50% (compared to DMU, WMU, and UWTOWTU respectively). Table 2 summarizes the NMBE median values for all algorithms and prediction windows.

Table 2. Median NMBE values for tested algorithms

	NMBE		
Prediction Algorithm	10:00am - 6:00pm (%)	12:00pm - 6:00pm (%)	1:00pm - 4:00pm (%)
DMU	16.9	17.5	17.3
DMPA	8.1	7.5	7.4
WMU	18.7	18.7	18.4
WMPA	10.3	7.4	7.0
UWTOWTU(7.0)	9.1	9.8	9.5

UWTOWTPA(7.0)	5.8	6.0	4.5
WTOWTU(70.10)	12.5	13.1	11.6
WTOWTU(70.14)	12.9	13.3	12.1

Conclusions

The research questions answered in this work are repeated below, with conclusions relating to each question.

Research question 1: How does the advanced M&V regression-based approach compare to the established averaging methods?

Industry-accepted baseline techniques and model-based approaches underpredicted peak period consumption for the selected test dataset. If this dataset and these methods had been used for real DR events, they would have significantly under-credited the DR load-reduction benefits. Median bias varied between algorithms, with the unweighted 7-day TOWT algorithm (with adjustment) having the lowest median bias value (Table 2).

We note that median bias values in this study are larger than many of the examples reported in prior literature—some of which are biased toward an underprediction, some toward an overprediction, and some near to zero. This may reinforce the notion that: (1) a high degree of customization is needed to identify an approach and adjustment method that will provide accurate predictions of peak building loads (e.g., based on climate, building loadshape characteristics, etc.), and (2) methods that work in one case are not assured to be generalizable.

Research question 2: Does the duration and timing of the DR event window have a significant impact on the prediction accuracy?

Given the observed similarity of median bias results across the three different time windows (Table 2), ranging from three to eight hours and with different start times, we conclude that duration and timing of the DR event did not have a significant impact on the prediction accuracy under this study.

Research question 3: Are there notable differences in the distribution of prediction accuracy results across a large population of meters when employing different baseline prediction methods?

Figure 4 shows there was significant overlap in the distributions across all algorithms tested, suggesting similar performance trends overall. Weather-matched algorithms exhibited the widest distribution in NMBE, and all algorithms saw an interquartile range exceeding 10%.

By definition, a peak day will see temperatures and loads outside the range observed prior to the event day, irrespective of the time window chosen on that particular day. Any baseline estimation approach will be limited in predicting consumption outside of the range of independent variables observed in the training period; by design, this study selected prediction days that represented the peak consumption for each meter, exacerbating this limitation. The study results, therefore, may represent the worst case in terms of error due to a lack of representative independent variable data during the training period.

Given that increasing levels of renewables are driving a need for building load flexibility in support of grid stability, these results highlight the opportunity to improve peak load prediction methods and to reduce the dependence on customized adjustments.

The test dataset used for this research was intentionally broad, covering a range of geographical regions to assess prediction robustness across a wide range of conditions. It is possible that a more intentionally curated dataset may allow for tailoring a more accurate prediction method limited to a narrower set of building typologies and climates. Identifying actual DR event days within a region and selecting commercial buildings' data from those days (for buildings that did not participate in the DR event) would be another potential approach to selecting test data.

Possible future research should explore different model types (e.g., machine learning, quantile regression) and/or assess the potential benefits from inclusion of different independent variables such as cooling load. Further study could also consider whether different algorithms might be matched to different buildings based on those buildings' loadshape characteristics (e.g., weather-dependency of load). Changing the pre-adjustment calculation and changing the cap value may be worth testing, though as noted above, there is a risk that this would result in an arbitrary calculation adjustment driven by a specific dataset and would not be generalizable across different regions, building types, etc. Further, if applied to DR programs, a higher adjustment cap would increase risk exposure for gaming.

Acknowledgements

This work was supported 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 for Measurement of Energy and Demand Savings." American Society of Heating, Refrigeration and Air Conditioning Engineers, Atlanta, GA.

Baechler, Michael C., Theresa L. Gilbride, Pam C. Cole, Marye G. Hefty, Kathi Ruiz. 2015. "Building America Best Practices Series Volume 7.3: Guide to Determining Climate Regions by County." Pacific Northwest National Laboratory. PNNL report number: PNNL-17211 Rev. 3. https://www.energy.gov/sites/prod/files/2015/10/f27/ba_climate_region_guide_7.3.pdf.

Bode, Josh, Adriana Ciccone. 2017. California ISO Baseline Accuracy Assessment. CAISO Baseline Accuracy Working Group.

CalTRACK. 2018. "CalTRACK Technical Documentation: Modeling Hourly Methods." Retrieved from: http://docs.caltrack.org/en/latest/methods.html#section-3-b-modeling-hourly-methods.

Chai, Yanxin, Yue Xiang, Junyong Liu, Chenghong Gu, Wentao Zhang, Weiting Xu. 2019. "Incentive-Based Demand Response Model for Maximizing Benefits of Electricity Retailers." Journal of Modern Power Systems and Clean Energy 2019; 7:1644-1650. DOI: https://doi.org/10.1007/s40565-019-0504-y.

Crowe, Eliot, Jessica Granderson, Samuel Fernandes. 2019. "From Theory to Practice: Lessons Learned from an Advanced M&V Commercial Pilot." Proceedings of the 2019 International Energy Program Evaluation Conference.

Eversource. 2020. "Earn Money & Save Energy: Earn incentives for helping reduce peak demand and carbon emissions." Eversource program marketing literature. Accessed December 21, 2020.

https://www.eversource.com/content/docs/default-source/save-money-energy/curtailment-demand-response.pdf?sfvrsn=8b3bc962_4.

Federal Energy Regulatory Commission. 2019. "2019 Assessment of Demand Response and Advanced Metering." FERC Staff Report. https://www.ferc.gov/sites/default/files/2020-04/DR-AM-Report2019_2.pdf.

Franconi, Ellen, Matt Gee, Miriam Goldberg, Jessica Granderson, Tim Guiterman, Michael Li, Brian A. Smith. "The Status and Promise of Advanced M&V: An Overview of M&V 2.0 Methods, Tools, and Applications." Rocky Mountain Institute, 2017 and Lawrence Berkeley National Laboratory, 2017. LBNL report number ##LBNL-1007125.

Goldberg, Miriam, Ken Agnew. 2013. "Measurement and Verification for Demand Response: Development of a Standard Baseline Calculation Protocol for Demand Response." National Forum on the National Action Plan on Demand Response: Measurement and Verification Working Group.

Granderson, Jessica, Samir Touzani, Eliot Crowe, Samuel Fernandes, Shankar Earni, Kaiyu Sun. 2019. "Realizing high-accuracy low-cost measurement and verification for deep cost savings." Final Project Report. DOI: https://dx.doi.org/10.20357/B7TS3G.

Granderson Jessica, Samir Touzani, Claudine Custodio, Michael Sohn, David Jump, Samuel Fernandes, 2016. "Accuracy of Automated Measurement and Verification (M&V) Techniques for Energy Savings in Commercial Buildings." Applied Energy, 173, pp.296-308.

Grimm, Clifford. 2008. "Evaluating Baselines for Demand Response Programs." AEIC Load Research Workshop.

Jarred, Michael W. 2014. "Delivering on the Promise of California's Demand Response Programs. Policy Matters." Policy Matters, June 2014. California Senate Office of Research. https://sor.senate.ca.gov/sites/sor.senate.ca.gov/files/SOR_Policy_Matters--Demand_Response.pdf.

KEMA, Inc. 2011. "PJM Empirical Analysis of Demand Response Baseline Methods White Paper." PJM Load Management Task Force.

Kiliccote, Sila, Mary Ann Piette, Johanna Mathieu, Kristen Parrish. 2010. "Findings from Seven Years of Field Performance Data for Automated Demand Response in Commercial Buildings." Proceedings of the 2010 ACEEE Summer Study on Energy Efficiency in Buildings. LBNL report number: LBNL-3643E.

Mathieu, Johanna L., Phillip Price, Sila Kiliccote, Mary Ann Piette. 2011. "Quantifying changes in building electricity use, with application to Demand Response." IEEE Transactions on Smart Grid 2:507-518.

Price, Phillip, Nathan Addy, Sila Kiliccote. 2015. "Predictability and Persistence of Demand Response Load Shed in Buildings." Lawrence Berkeley National Laboratory. LBNL report number: LBNL-187399.

Xenergy, Inc. 2002. "Protocol Development for Demand Response Calculation: Draft Findings and Recommendations." California Energy Commission. http://www.calmac.org/publications/2002-08-

02_XENERGY_REPORT.pdf.