Title

Assessment of Model-Based peak electric consumption prediction for commercial buildings

Authors

Granderson, Jessica
Sharma, Mrinalini
Crowe, Eliot
et al.

Publication Date

2021-08-01

DOI

10.1016/j.enbuild.2021.111031

Peer reviewed

Energy & Buildings 245 (2021) 111031

Contents lists available at ScienceDirect

Energy & Buildings

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

Assessment of Model-Based peak electric consumption prediction for commercial buildings

Jessica Granderson [a], Mrinalini Sharma [b], Eliot Crowe [a,*], David Jump [b], Samuel Fernandes [a], Samir Touzani [a], Devan Johnson [b]

[a] *Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA*
[b] *kW Engineering, 287 17th Street, Suite 300, Oakland, CA 94612, USA*

ARTICLE INFO

Article history:
Received 14 January 2021
Revised 2 April 2021
Accepted 13 April 2021
Available online 19 April 2021

Keywords:
Demand Response (DR)
Demand side management (DSM)
Energy efficiency (EE)
Measurement and verification (M&V)

ABSTRACT

Utility programs have successfully delivered energy efficiency for decades. Today, increasing emphasis is being placed on demand response (DR) programs that incentivize customers to reduce, or "shed" electric load during grid peak periods. The most common methods used to predict building peaks and quantify DR load reductions rely on simple averaging algorithms using hourly load and temperature data from the days preceding the DR event. In contrast, regression-based algorithms have been used for decades to quantify annual energy efficiency savings. The availability of smart meter data has enabled application of hourly regressions for more accurate energy savings estimation, often referred to as "advanced measurement and verification (M&V)." This project explored whether advanced M&V regression approaches offer improvements over simpler averaging approaches for peak load prediction in commercial buildings.

We present evaluation results for eight algorithms (based on three baseline modeling approaches). The findings show that all algorithms underpredicted consumption across 453 meters and over 1,100 peak load days. Median bias values varied between 4.5 and 18.7 percent, indicating that the methods evaluated would tend to understate achieved load reductions in DR applications for these buildings. The regression methods did not offer a notable advantage over the commonly used averaging methods.

1. Introduction

The concept of demand response (DR) can be traced to the beginnings of the U.S. electric power industry (circa early- to mid-1890s), where system engineers and utility executives debated the optimal pricing regime for this newfound service [6]. In recent decades, DR programs have evolved around two basic approaches: rate-based and incentive-based [8]. The Federal Energy Regulatory Commission (FERC) defines DR as "Changes in electric usage by end-use customers from their normal consumption patterns in response to changes in the price of electricity over time, or to incentive payments designed to induce lower electricity use at times of high wholesale market prices or when system reliability is jeopardized" [3]. 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 percent 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") [19]. As of 2017, almost 19 million utility customers were enrolled in DR programs across the United States [13].

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 [12].

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). 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

[*] Corresponding author.
E-mail address: ecrowe@lbl.gov (E. Crowe).

https://doi.org/10.1016/j.enbuild.2021.111031
0378-7788/© 2021 The Authors. Published by Elsevier B.V.

Nomenclature

DR	Demand response
TOWT	Time-of-Week-and-Temperature energy model
DM	Day-matching
WM	Weather-matching

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.

Much prior work has been conducted to calculate counterfactual predictions; for example, considering (1) the duration of the time window from which baseline days are selected; (2) criteria for selecting specific days within that pre-event time period; and (3) the calculation approach and any related adjustments. The most relevant examples of this prior work are summarized below.

A California-based 2017 study [4] 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 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. The rationale for adjustment is that the difference in load during the pre-/post-event period can be treated as measurement error, and the adjustment process reduces that error. In this study the adjustments were capped at ± 20 percent for the day-matching methods and ± 40 percent for the weather-matching methods. 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 [21]. 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. 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 over-counting), (3) make adjustments to program rules, and (4) treat the DR program as an iterative process, adapting the program M&V approach based on ongoing results and the customer mix [15]. A 2009 study analyzed nonresidential building baseline models, classified buildings into four types with different degrees of load variability and weather sensitivity, and found that the accuracy of baseline load models can be improved substantially by applying pre-event adjustments to baseline predictions [9].

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 [27].

Similar to [27], a 2008 study also noted a need for DR baseline methods to minimize the risk of gaming [18], and found that methods using multiple pre-event days reduced the risk of gaming (e.g., with a short notice period prior to a DR event, a customer could not deliberately inflate their consumption for 10 days prior to the event). However, similar to most studies, they found that an adjustment to the baseline calculation is needed to most accurately estimate actual customer usage.

While DR assessments have always required interval meter data, energy efficiency applications have not. However advanced measurement and verification (M&V) methods have emerged over the past decade, employing hourly or subhourly data and sophisticated modeling approaches to quantify energy efficiency savings with a high degree of accuracy [14]. The "time of week and temperature" (TOWT) model is a piecewise linear regression that has been well documented in the literature [17,24]. The TOWT model and its variants also have been incorporated into utility program efficiency M&V and industry tools as an accepted method [5,10,16]. Some of the first uses of this model targeted DR applications [22,24,25]. [25] assessed the predictive accuracy of a more complex variation of the TOWT model, with a custom adjustment

[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).

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 percent ("baseline percent error," where a positive value indicates the predictions were higher than actual consumption), compared to 6 percent for a day-matched 10-day algorithm and 5 percent 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. For example, [11] focuses on predicting cooling load using artificial neural networks [11], and [20] focuses on predicting heating load using similar methods [20]. [23] is one of many papers which tests electricity energy model predictive accuracy over a longer time period (weeks), and also on a limited dataset from two buildings [23]. Chae et al. [7] is an example where the prediction window (day-ahead) may align with DR baseline applications, but the proposed method employs HVAC temperature setpoints as an input [7], making it impractical for scaled deployment through DR programs (i.e., HVAC temperature setpoints are not easily obtained at scale, and may or may not be affected by the demand reduction strategies deployed in a given building). [26] is another example of a sophisticated modeling technique that uses 11 input features including occupancy [26], that shows potential for accurate prediction but is not easily scalable due to model input data availability.

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?

Section 2 of this paper describes the methodology underlying the study, Section 3 summarizes the findings, and Section 4 provides a discussion of the results. Section 5 provides conclusions and ideas for future work.

2. Method

The evaluation of DR baseline predictive accuracy presented in this paper is based on a five–step assessment process:

1. Collate a dataset of hourly load and ambient temperature data for commercial buildings' meters with no known efficiency improvements or DR events.
2. 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.
3. 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.
4. Repeat the steps above for all meters in the dataset, and quantify the distribution of error metrics for each algorithm.
5. Compare the distributions and median error metrics for each algorithm.

This study was targeted at commercial buildings; DR programs target commercial, residential, and industrial sectors, but only commercial buildings' data was available for this study. 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 to allow for the fact that DR event windows may occur during different time windows, depending on region, generation mix, and weather conditions.

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

2.1. Data preparation

The test dataset comprised 12 months of hourly electric consumption (kilowatt-hours, kWh) and corresponding 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 came from buildings located in three U.S. Building America climate zones [2]: Marine, Cold, and Mixed-Humid. The test dataset was intentionally diverse in terms of region, consumption, and property type, to allow for assessment of peak prediction algorithms across a diverse set of conditions. All data were cleaned of obvious erroneous values, such as temperature values below −34.5 °C (-30°F) and above 65.5 °C (150°F).

For each meter, the ten non-holiday, non-weekend days with the highest maximum daily load were identified. These were selected as candidates for days on which peak loads would be predicted for this study. Any candidate day that did not have a sufficient history of data to satisfy *all* of the baseline methods tested (with baseline time periods ranging from 10 to 90 days) was excluded. The result was 1,104 prediction days that were used in the analysis, a sufficiently large quantity to determine overall performance and variability of prediction results.

2.2. Peak prediction algorithms

Three peak prediction algorithms were assessed in this study: two averaging algorithms and a regression method. Each of these methods, and the variants tested, are described below.

2.2.1. Averaging methods

Two averaging methods were selected based on the best results reported in [4]:

- *Day-Matching:* Baseline data are drawn from the 10 working days immediately prior to the event day. For each hour of the event day, the corresponding hours from the baseline data are averaged to calculate hourly predictions for the event window.
- *Weather-Matching:* Baseline data are drawn from the 4 days out of the 90 days prior to the event, with maximum temperature closest to the maximum temperature of the event day. For each hour of the event, the corresponding hours from the baseline data are averaged to calculate hourly predictions for the event window.

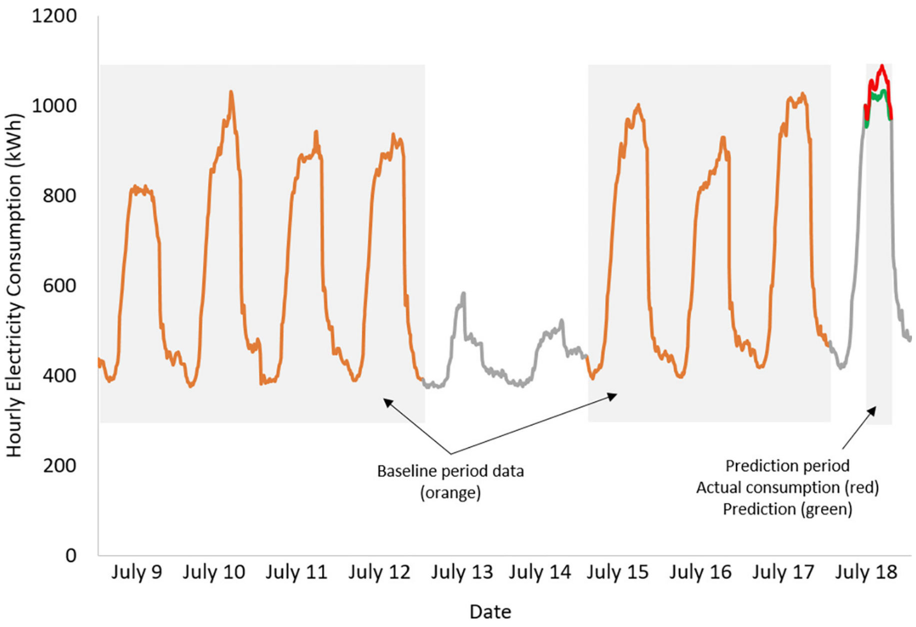

Fig. 1. Example plot showing the prediction period (10:00 am to 6:00 pm on July 18) and baseline data used for one prediction algorithm studied.

The literature review did not conclude that these two averaging methods are objectively the best for DR applications. They were selected for this study as two contrasting options that had been shown to perform well and are in current use.

2.2.2. Regression method

The regression method selected for this study was the TOWT 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 [17] the TOWT model was shown to be highly accurate at predicting annual consumption, equaling or outperforming other M&V industry standard models. The TOWT model uses time of the week and the outside air temperature as input variables, and it 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). Three variants of the TOWT model were tested under this study:

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

2.2.3. Adjustments for Day-of-Event conditions

As reported in prior literature, adjustment methods have been developed to account for weather impacts to load on peak days. These methods are based on the observed load before or after the event window. In this work, the adjustment approach documented in [4] was used.

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).

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

Adjustment ratio caps applied in this work followed the recommendations in [4]:

- Weather-Matching: +/- 40 percent (also applied to TOWT regression)
- Day-Matching: +/- 20 percent

Table 1 lists each of the algorithms and variants tested.

2.3. Assessment metrics

Many possible metrics can be used to quantify the error of model predictions. Different metrics provide different insights into aspects of performance. In prior work to assess accuracy of advanced M&V models, normalized mean bias error (NMBE) and the coefficient of variation of the root mean squared error (CV [RMSE]) were used [17]. NMBE and CV(RMSE) are also familiar to practitioners, and are prominent in resources such as ASHRAE Guideline 14 [1]. NBME and CV(RMSE) are defined in Equations (2) and (3) below, where y_i is the actual metered value, $\hat{y}_i$ is the predicted value, $\bar{y}$ is the average of the y_i, and N is the total number of data points.

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.

$$NMBE = \frac{\frac{1}{N}\Sigma_i^N\left(y_i - \widehat{y}_i\right)}{\bar{y}} \times 100 \quad (2)$$

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

For this study the metrics were being calculated based on model predictions of data that were not used in the model creation, in a process known as *cross-validation* or *out-of-sample testing.* NMBE and CV(RMSE) values closer to zero indicate more accurate predictions. For NMBE, bias may be positive or negative, with positive values indicating underprediction (i.e., predicted values are lower than actual values). Both metrics are applied to assess the accuracy of each individual model's predictions, as opposed to assessing variability across the whole population.

3. Results

Fig. 2 shows a single prediction window for each algorithm that was evaluated. The plot shows the actual hourly meter readings and the studied algorithms' predictions for the hours between 10:00am and 6:00 pm on a single day. This provides a visual example as context for the results that follow, that summarize predictive accuracy across all the models and across each of the 1,104 prediction days.

Figs. 3 and 4 show the distribution of NMBE and CV(RMSE) results, respectively, 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.

The results indicate that there was not a significant difference in predictive accuracy for the three prediction windows, and there was no consistent pattern in terms of which prediction window saw the highest or lowest median values. Fig. 3 also illustrates the significant overlap in distribution between all the algorithms tested. The median CV(RMSE) improved as the start of the prediction window moved later and the prediction window was shorter, though the improvements were modest, e.g., the UWTOWTU(7.0) model saw median CV(RMSE) values of 17.6 percent, 16.9 percent, and 14.1 percent for the three prediction windows. Similar to the NMBE results, CV(RMSE) results did not see a large shift in distribution across the three prediction windows, and there was a high degree of overlap in distribution between the algorithms.

Based on distribution in values for the CV(RMSE) metric, day-matching with adjustment (DMPA) performed best (i.e., had the tightest distribution and lowest median), while weather-matching without adjustment (WMU) performed worst (i.e., had the widest distribution and highest median). There was relatively little difference between the other algorithms tested, with a high degree of overlap between their distributions.

For the NMBE metric, the results also reflect a wide distribution for all methods (with the weather-matching algorithm having the widest distribution), with significant overlap between algorithms. All values were biased in a positive direction, meaning that each algorithm tested underestimated the load during peak hours. The lowest median bias (NMBE 4.5 percent) was observed for the unweighted, adjusted TOWT regression model with the 1:00 pm–4:00 pm prediction window; however, this median value is still considered high. Additionally, the results show that the application of adjustments had a significant effect on NMBE; in six out of nine cases the adjustment reduced the median NMBE value by more than half.

Table 2 summarizes the NMBE and CV(RMSE) median values for all the algorithms and prediction windows.

For additional perspective into relative algorithm performance across both error metrics, Fig. 5 shows the median NMBE and CV (RMSE) results for the 12:00 pm–6:00 pm event window as a scatter plot. The DMPA algorithm combines the lowest median CV (RMSE) and the near-lowest NMBE. It should be noted that ideal NMBE values would be zero, whereas there is no industry-accepted target value for CV(RMSE) in the context of DR timescales; for reference, however, a CV(RMSE) target of less than 25% is typically desired for whole-year regression models. Further, the relative level of importance of low bias and low variability is subjective.

4. Discussion

The baseline prediction methods considered in this analysis underpredicted the energy consumption for the prediction window

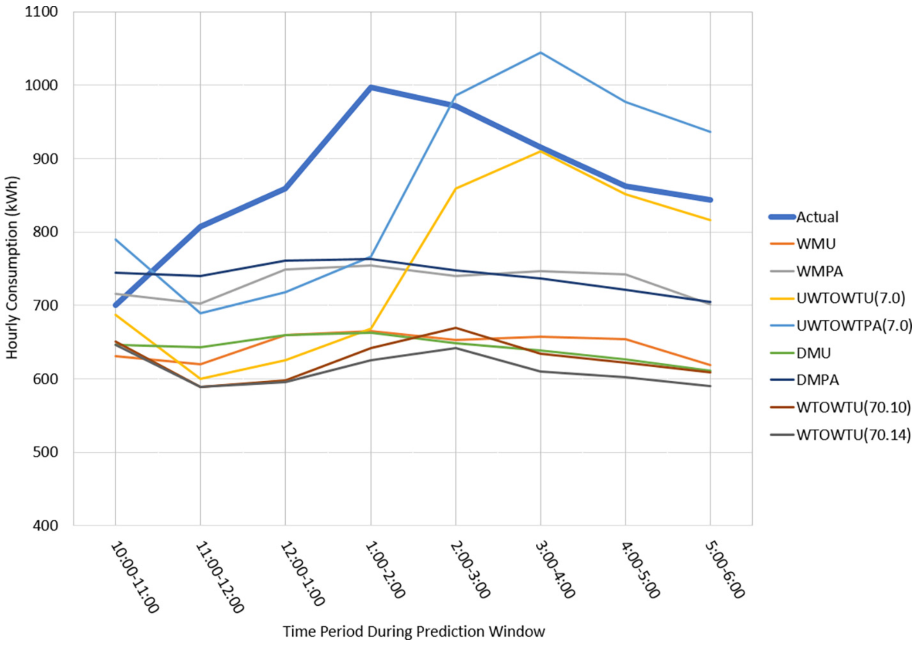

Fig. 2. Example prediction window (10:00am to 6:00 pm), illustrating the predictions from each of the tested algorithms compared with actual consumption.

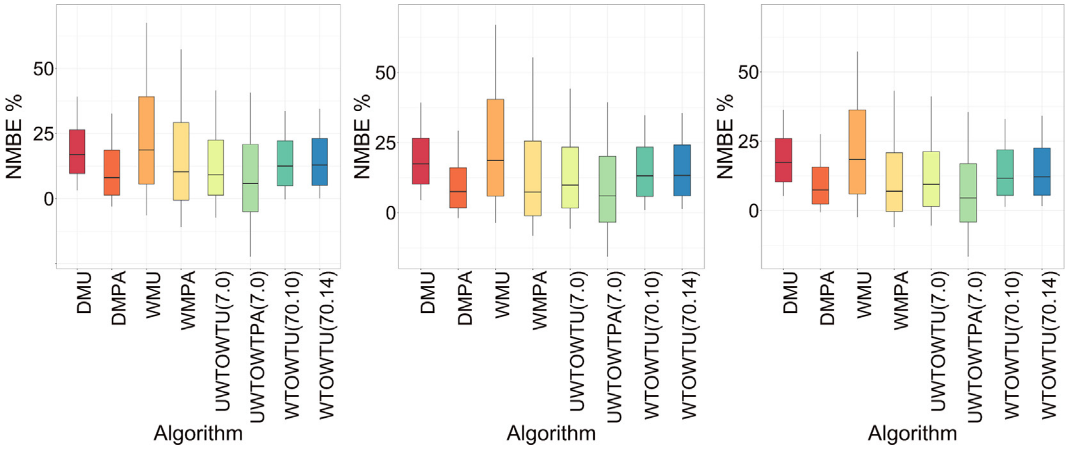

Fig. 3. NMBE results distribution for 10:00 am–6:00 pm (left), 12:00 pm–6:00 pm (center), and 1:00 pm–4:00 pm prediction windows.

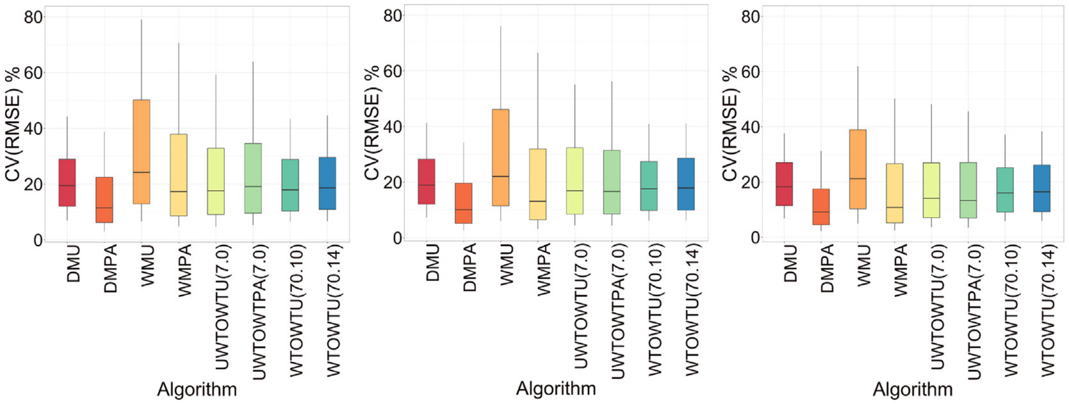

Fig. 4. CV(RMSE) results distribution for 10:00 am–6:00 pm (left), 12:00 pm–6:00 pm (center), and 1:00 pm–4:00 pm prediction windows.

Table 2
Median values for assessment metrics.

	NMBE			CV(RMSE)		
Prediction algorithm	10:00am–6:00 pm (%)	12:00 pm–6:00 pm (%)	1:00 pm–4:00 pm (%)	10:00am–6:00 pm (%)	12:00 pm–6:00 pm (%)	1:00 pm–4:00 pm (%)
DMU	16.9	17.5	17.3	19.5	18.9	18.3
DMPA	8.1	7.5	7.4	11.5	10.1	9.1
WMU	18.7	18.7	18.4	24.2	22.0	21.3
WMPA	10.3	7.4	7.0	17.4	13.1	10.8
UWTOWTU (7.0)	9.1	9.8	9.5	17.6	16.9	14.1
UWTOWTPA (7.0)	5.8	6.0	4.5	19.1	16.7	13.3
WTOWTU (70.10)	12.5	13.1	11.6	17.9	17.6	16.0
WTOWTU (70.14)	12.9	13.3	12.1	18.7	17.9	16.4

period in the majority of cases (i.e., NMBE was greater than zero for approximately 85 percent of prediction windows). Median NMBE values ranged from 4.5 percent to 18.7 percent, a significant deviation that would result in undercounting DR load reductions if applied in a program context. This is particularly significant given that a 5 percent understatement of counterfactual energy consumption would constitute a much larger understatement of any claimed DR load reduction. Further, the distribution of NMBE results was very wide in all cases (a 20 percent or greater range between the 25th and 75th percentiles). While there was some variation in the results produced by different algorithms, and the bias was partially mitigated by adjustments where applied, these general observations hold for all eight methods tested, across all three event time windows.

It is interesting to 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. For example, [27] and [4] assessed many algorithms and reported bias within ± 2 percent for several algorithms. This is in part due to differences in datasets and methodology. For example, some studies used data from a less diverse set of climates or building types, some used smaller datasets, and some considered average loads across several peak hours. However, it also indicates that: (1) a high degree of customization

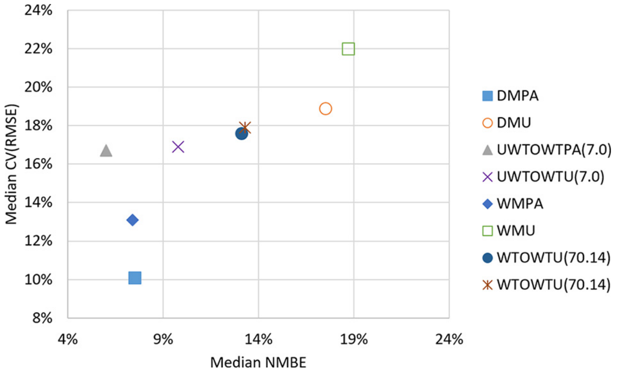

Fig. 5. Median NMBE and CV(RMSE) values for each tested algorithm (prediction window 12:00 pm to 6:00 pm).

is needed to identify an approach and adjustment method that that will provide accurate predictions of peak building loads, and (2) methods that work in one case are not assured to be generalizable.

No significant differences were found in the performance of the algorithms when applied to the three event/prediction time windows. The possible reasons for this were not studied, though it may the case that day effects are more dominant that within-day effects. Further, 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 (and perhaps explaining why this study reported higher bias relative to prior studies). In practice, the system peaks that drive DR dispatches typically occur on hot days when commercial buildings' cooling load is high, but it is not guaranteed that the system peak days will correspond with all buildings' peak consumption days. 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. Further, the impact of selecting peak consumption days as prediction periods may affect different prediction algorithms to different extents.

The similarity of results for different time windows may also be an indicator of robustness of the methods across a range of conditions, which may be useful when considering that the timing of generation system peaks is changing as more renewables are integrated with the grid. Future work could look deeper into the bias for each hour of the targeted time window, to better understand predictability characteristics across the range of temperatures and load magnitudes.

Baseline adjustments are intended to account for a lack of similar conditions in the baseline period (specifically, that the DR event day is hotter than preceding days), and have been traditionally applied to averaging methods. In this analysis, we observed that the unadjusted, unweighted TOWT (UTOWTU) performs at bias levels only slightly higher than the adjusted averaging methods (DMPA and WMPA). The adjustment caps detailed earlier in the document are used to limit the potential for manipulation of loads to influence the baseline. The event window of 10:00 am–6:00 pm pushes the adjustment hours to pre-startup conditions (6:00 am–8:00 am), and as such, using these (often unoccupied) hours to adjust the prediction offered little benefit. Increasing the pre-adjustment cap value would have improved the NMBE values under this test, though not significantly (i.e., the cap was only applied in approximately 15 to 30 percent of cases anyway). 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. Another option that has been implemented in prior work is to apply an adjustment based on conditions immediately after the event window, instead of, or in addition to, conditions prior to the event, or even on surrounding days. For example, [4] tested the use of pre- and post-event hours and found a slight improvement over using pre-event hours only.

The unweighted TOWT algorithm was run with a training period of seven days. Since the prediction events and associated baseline groups excluded weekends, this arrangement created uneven training periods for the algorithm—two of the five weekdays had double the representation in the training dataset. Using 5 or 10 *weekdays* for the TOWT algorithm would address this imbalance, though we did not study the effect of the imbalance on the load predictions.

The test dataset used for this research was intentionally broad, covering a range of geographical regions and not limiting the commercial sectors, in order to assess prediction robustness across a wide range of conditions. It is possible that a more intentionally curated dataset may allow for deeper study of the relationship of predictive accuracy and building loadshape, and possibly the tailoring of a more accurate prediction method limited to a narrower set of building typologies and climates. The approach of selecting days on which peak loads occurred for meters may also be studied further; it is assumed that those would be the days most likely to experience a DR event, but it is possible that identifying peak temperatures or other factors may be more appropriate selection criteria. For example, 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.

5. Conclusions and future work

For the subject dataset of 453 meters (divided into over 1,100 prediction days), industry-accepted baseline techniques and model-based approaches underpredicted consumption for the target time windows (10:00 am to 6:00 pm, 12:00 pm to 6:00 pm, and 1:00 pm to 4:00 pm). If this dataset and these methods had been used for real DR events, they would have undercredited the DR load-reduction benefits by 4.5 to 18.7 percent. There were differences in median bias between algorithms, but all consistently underpredicted peak period consumption, and there was significant overlap in the distributions across all algorithms tested, suggesting similar performance overall.

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. 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).

CRediT authorship contribution statement

Jessica Granderson: Conceptualization, Methodology, Supervision, Funding acquisition. **Mrinalini Sharma:** Methodology, Software, Formal analysis, Data curation, Visualization. **Eliot Crowe:** Methodology, Writing - original draft, Visualization, Project administration. **David Jump:** Methodology, Formal analysis. **Samuel Fernandes:** Software, Data curation, Visualization. **Samir Touzani:** Methodology. **Devan Johnson:** Methodology, Formal analysis.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

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

[1] ASHRAE, ASHRAE Guideline 14–2014 for Measurement of Energy and Demand Savings, American Society of Heating, Refrigeration and Air Conditioning Engineers, Atlanta, GA, 2014.

[2] Michael C. Baechler, 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.

[3] V.S.K. Balijepalli, V.P. Murthy, S.A. Khaparde, R.M. Shereef, 2011. ''Review of Demand Response under Smart Grid Paradigm.'' ISGT2011-India, Kollam, Kerala, 2011, pp. 236-243, doi: 10.1109/ISET-India.2011.6145388.

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

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

[6] Cappers Peter, Charles Goldman, David Kathan. Demand response in U.S. Electricity markets: empirical evidence, Energy 2010(4) 2010 1526–1535. DOI: https://doi.org/10.1016/j.energy.2009.06.029.

[7] Young Tae Chae, Raya Horesh, Youngdeok Hwang, Young M. Lee, Artificial neural network model for forecasting sub-hourly electricity usage in commercial buildings, Energy Build. 2016 (111) (2016) 184–194, https://doi.org/10.1016/j.enbuild.2015.11.045.

[8] Yanxin Chai, Yue Xiang, Junyong Liu, Gu. Chenghong, Wentao Zhang, Xu. Weiting, Incentive-based demand response model for maximizing benefits of electricity retailers, J. Modern Power Syst. Clean Energy 2019 (7) (2019) 1644–1650, https://doi.org/10.1007/s40565-019-0504-y.

[9] Katie Coughlin, Mary Ann Piette, Charles Goldman, Sila Kiliccote, Statistical analysis of baseline load models for non-residential buildings, Energy Build. 2009 (41) (2009) 374–381, https://doi.org/10.1016/j.enbuild.2008.11.002.

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

[11] Chirag Deb, Lee Siew Eang, Junjing Yang, Mattheos Santamouris, Forecasting diurnal cooling energy load for institutional buildings using Artificial Neural Networks, Energy Build. 2016 (121) (2016) 284–297, https://doi.org/10.1016/j.enbuild.2015.12.050.

[12] 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.

[13] 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.

[14] 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.

[15] Miriam Goldberg, 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.

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

[17] Granderson Jessica, Samir Touzani, Claudine Custodio, Michael Sohn, David Jump, Samuel Fernandes, Accuracy of automated measurement and verification (M&V) techniques for energy savings in commercial buildings, Appl. Energy 173 (2016) 296–308.

[18] Clifford Grimm. 2008. ''Evaluating Baselines for Demand Response Programs.'' AEIC Load Research Workshop.

[19] Michael W. Jarred, 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.

[20] Radiša Ž. Jovanović, Aleksandra A. Sretenović, Branislav D. Živković, Ensemble of various neural networks for prediction of heating energy consumption, Energy Build. 2015 (94) (2015) 189–199, https://doi.org/10.1016/j.enbuild.2015.02.052.

[21] KEMA, Inc. 2011. ''PJM Empirical Analysis of Demand Response Baseline Methods White Paper.'' PJM Load Management Task Force.

[22] Sila Kiliccote, Mary Ann Piette, Johanna Mathieu, Kristen Parrish, 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, 2010.

[23] Kangji Li, Su. Hongye, Jian Chu, Forecasting building energy consumption using neural networks and hybrid neuro-fuzzy system: a comparative study, Energy Build. 2011 (43) (2011) 2893–2899, https://doi.org/10.1016/j.enbuild.2011.07.010.

[24] Johanna L. Mathieu, Phillip N. Price, Sila Kiliccote, Mary Ann Piette, Quantifying changes in building electricity use, with application to demand response, IEEE Trans. Smart Grid 2 (3) (2011) 507–518.

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

[26] Zeyu Wang, Yueren Wang, Ravi S. Srinivasan, A novel ensemble learning approach to support building energy use prediction, Energy Build. 2018 (159) (2018) 109–122, https://doi.org/10.1016/j.enbuild.2017.10.085.

[27] Inc Xenergy, Protocol development for demand response calculation: draft findings and recommendations, California Energy Commission (2002). http://www.calmac.org/publications/2002-08-02_XENERGY_REPORT.pdf.