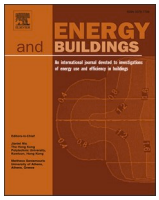
Energy & Buildings 295 (2023) 113297

Contents lists available at ScienceDirect

Energy & Buildings

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



Accuracy of hourly energy predictions for demand flexibility applications

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

Keywords:
Demand Flexibility (DF)
Demand Response (DR)
Advanced metering infrastructure (AMI)
Demand side management (DSM)
Energy efficiency (EE)
Measurement and verification (M&V)
Load shift (LS)
Load shed
Time-sensitive
Load prediction

ABSTRACT

Decarbonization goals in the United States electricity sector are increasing the levels of renewable energy generation in the electricity supply system, and are driving increased attention to building electrification, which will increase the magnitude and shift the timing of the electricity system peak. These changes are motivating new approaches to coordinate building electricity demand with low-carbon renewable generation, elevating the importance of demand flexibility (DF) in buildings and the need to quantify the temporal impacts of DF. In this paper, we first characterize the hourly predictive accuracy of six commonly used baseline models in an application context of quantifying building-level load shift. Our analysis revealed insights such as hours of the day (afternoons), periods of the week (weekends), and seasons (summer) that were predicted with more accuracy than other time periods. In addition, the analysis showed tendencies toward overprediction or underprediction of load. Secondly, we provide the first published investigation of baseline erosion from repeated dispatch of building load shifting. We observed that as the baseline period is pushed back further from the prediction day, the distribution of errors across baseline model predictions increases, with notable inflection points near the three-week erosion point for two of the three models.

1. Introduction

Electricity sector decarbonization goals and renewable portfolio standards are increasing the levels of renewable energy generation into the electricity supply system. In 2021, renewable energy sources accounted for 19.8% of U.S. electricity generation [8] and globally, the roadmap of the International Agency for Renewable Energy (IRENA) forecasts a share of renewable energy beyond 30% by 2030 [10]. At the same time, building sector decarbonization goals are driving increased attention to electrification, which will increase load, and shift the timing of the system peak. Similar impacts are anticipated from vehicle electrification. These changes are motivating new approaches to coordinate building electricity demand with low-carbon renewable generation, elevating the importance and value of demand flexibility (DF) in buildings. Building-provided grid services can be delivered through various modes of DF such as load shed, load shift, and load modulation, enabled by advanced controls, communications, and analytics [5]. Excerpted from [5], Table 1 summarizes examples of building DF modes mapped to associated grid services, with a description of the change in building operation and key characteristics such as duration of change, load change, response time and event frequency. Historically, load shed has been deployed through demand response programs (DR) designed to alleviate peak loads on the grid. An emerging and growing concern is the curtailment of renewables that occurs when supply exceeds net demand. For instance, in California, solar curtailment has increased in recent years to over 5% of utility scale solar generation in 2020 [9], and across all Independent System Operators (ISOs), wind curtailment is growing and reached 4.8% in 2021 [11].

Over the past decade, advanced measurement and verification (M&V) approaches have emerged employing hourly or sub-hourly data and sophisticated modeling approaches to quantify annual energy efficiency savings with a high degree of accuracy [7]. Assessing DF in buildings and appropriately valuing energy efficiency requires M&V methods and baseline models that capture changes in load at different times of the day. Mims et al. 2019 highlighted and reviewed the time-varying value of energy efficiency savings to the overall power system, and found that applying the time-sensitive value of efficiency can lead to planning or programs that more accurately value efficiency savings and identify the most valuable savings [30]. This is because the magnitude of savings that are generated by a given measure varies over the course of the day and the course of a year, as does the avoided cost of

* Corresponding author.
E-mail address: sgfernandes@lbl.gov (S. Fernandes).

https://doi.org/10.1016/j.enbuild.2023.113297
Received 27 January 2023; Received in revised form 27 May 2023; Accepted 20 June 2023
Available online 23 June 2023

Nomenclature

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

generating, transmitting, and distributing electricity. For example, [31] found a factor of 1.4–1.5 increase in the value of residential air conditioning measures, when using hourly savings shapes instead of non-time sensitive savings totals [31]. Similarly, quantifying DF requires knowing how much load changed during specific hours of the day and seasons or times of the year. This marks a difference from traditional quantification of efficiency that uses total aggregate saving, usually over an annual period as the base measure [1,13].

Prior work has demonstrated and applied methods to assess the accuracy of baseline models used for efficiency evaluation as well as those used for evaluation of peak load reductions. A study commissioned by the PJM Load Management Task Force (Kema [12,3] assessed several DR baseline approaches (including averaging and regression approaches) and found that the predictive accuracy can vary based on weather-responsiveness of load and the timing and season of the event window. A California-based 2017 study [6] assessed the accuracy and precision of different baseline strategies. The study found that multiple baseline approaches can deliver sufficiently unbiased and precise baselines for pooled aggregates of buildings, including weather-matched and day matched algorithms. The study also found that overall weather-matched methods typically outperformed day-matched methods. [32] characterized the performance of common baseline load models including day-matched and weather-matched approaches and noted that day-matched and weather-matched approaches are largely dependent on the characteristics of a smaller number of admissible days available for matching [32]. While evaluating the accuracy of day-matched and weather-matched approaches, [33] found that these methods performed well. However, they did find it difficult to find equivalent days for weather matching and highlighted that their accuracy was significantly improved by an adjustment applied to certain hours of the day. [4] and EnerNOC 2011 analyzed the accuracy and bias of DR baselines, and suggested different adjustments to improve baseline accuracy [4], Enernoc [12,3]. Granderson et al. [14], 2015b, 2016 presented consensus-based metrics and applied them to assess baseline predictive accuracy for different M&V models across an entire year of prediction period data to quantify annual savings impacts [14,15]. This work established the predictive accuracy of fully automated M&V methods in the context of quantifying changes in total annual energy use. [17] extended the application of M&V methods to quantify the hourly impacts of an entire DSM portfolio on the distribution grid, using models of aggregated consumption at the substation and feeder level [17]. The work provided a methodological and modeling foundation to connect efficiency programs with distribution grid-level planning. It also demonstrated the calculation of hourly savings shapes by season, but did not assess the hourly predictive accuracy of the applied methods. [18] highlight that while progress has been made to define baselines for shed and shift working in conjunction, there is no consensus on the best baseline approaches for these combined services [18]. They also draw attention to the fact that baseline dependent M&V methods will fail when an insufficient number of days meet baseline criteria, such as might occur as DF becomes more common, and more frequent and varied strategies are employed at a given site. [19] explored whether advanced M&V regression methods offer improvements over simpler averaging methods for prediction of the peak load in commercial buildings and found that the 8 tested baseline prediction algorithms underpredicted peak period consumption [19].

A gap in the body of existing work is that there is limited research on the accuracy of building-level hourly electricity use predictions and associated hourly quantification of changes in meter-level energy use or demand, which is more relevant for year-round DF applications. There is also a lack of published research that characterizes the impact of DF event frequency on baseline predictive accuracy. Finally, while the literature does cover accuracy in quantifying load shed at times of peak consumption, load shift time periods of interest have not been adequately investigated (i.e., load shift may occur over different periods of day and season compared to established load shed periods). This paper aims to address those gaps by answering the following questions:

- To what extent does prediction error vary by hour of day and by season, and what does that say about our ability to accurately quantify load shift at different times of the day and year?
- Are there consistent biases in hourly predictions that will impact quantification of load shift?
- To what extent does baseline erosion affect baseline prediction accuracy, and how much do these effects vary across different baseline algorithm forms?

The paper makes two key contributions. Firstly, it characterizes the hourly predictive accuracy of six commonly used baseline models in an application context of quantifying building-level load shift. Secondly, it provides the first published investigation of baseline erosion from repeated dispatch of building load shifting.

2. Method

This section discusses the dataset and models that were used in the study, the methods that were applied to assess hourly predictive accuracy and the degradation of predictive accuracy due to frequently implemented DF.

2.1. Predictive accuracy assessment: dataset, models, method, and metric

The test dataset that was used to assess the accuracy of hourly load predictions comprised metered data from 120 commercial buildings drawn from an existing dataset available to the researchers. There were no known energy efficiency projects or demand response (DR) events that had occurred in these buildings during the 24-months that the data covered. The data came from buildings located in two ASHRAE climate zones - marine, and mixed-humid [2]. The test dataset was intentionally diverse in terms of region and consumption, to expose the baseline

Table 1
Building demand flexibility modes and associated grid services [5].

DF mode	Grid service	Description of building change	Key characteristics [typical duration, load change, response time, event frequency]
Load Shed	Contingency reserves	Load reduction for a short time to make up for a shortfall in generation	Up to 1 hr, Short term decrease, <15 min, 20 times per year
Load Shift	Avoid renewable curtailment	Load shifting to increase energy consumption at times of excess renewable output	2–4-hour, short term shift, N/A, Daily
Load Modulation	Ramping	Load modulation to offset short-term variable renewable generation output changes	Seconds to minutes, rapid increase/decrease, seconds to minutes, continuous

models to a range of conditions. Data cleaning was mainly performed using two functions to clean erroneous temperature and load values. The clean temperature function removed observations that have temperature values higher than 54 degree centigrade or lower than −34 degree centigrade. A small number of data points were removed using this function (n = 153). The clean eload function removed observations that had negative load values. Again, a small number of data points were removed using this function (n = 17). There were no missing values in the dataset used as that was a pre-condition for dataset selection. Given the small number of data points removed in the data cleaning process, we do not anticipate this had any impact on the forecast performance.

Six different baseline models representing two different approaches - averaging and regression - were assessed in this study. Averaging approaches included day-matching and weather-matching and regression approaches included the Time of week and temperature (TOWT) and Gradient boosting machine (GBM) and their variants. A brief description of the different models used is provided below.

3. Averaging approaches

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

4. Regression approaches

Time of week and temperature (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 [21]. Two variants of the TOWT algorithm were tested under this study:

- Time of week and temperature with 7-day baseline (TOWT7): TOWT7 uses 7 baseline days prior to the day being predicted.
- Time of week and temperature with 70-day baseline (TOWT70): TOWT70 uses 70 baseline days prior to the day being predicted.
- Gradient Boosting Machine (GBM): The GBM baseline is an ensemble trees-based machine learning method [22]. The GBM generated a model of the energy consumption using time and temperature as independent variables.
- GBM Temperature and Energy Weighted (GBMTe50): This is a variant of the GBM where all the data in the training period are leveraged, proportional to a weighted calculation involving a sample's temperature and energy consumption. More weight is assigned to data with high temperature and high load.

These models were selected because prior studies have shown that they offer good predictive performance for building electricity use [21,16,22]. Model predictive accuracy was assessed using a variant of the published method in Granderson et al. 2015a. The steps involved in the procedure were:

i) Baseline model construction: Split the test dataset from each building into model training and model prediction periods. (The six baseline models' specifications from the literature define different amounts of baseline data). Since occupancy and consumption patterns are different for weekdays and weekends, baseline model for these days are constructed from their respective day types.

ii) Generate predictions: For each trained baseline model, generate predictions for each hour of the day, for one year of data, i.e., an entire year.

iii) Assess predictive accuracy: Compare the predictions to the data from the prediction period, and compute statistical performance metrics for every hour in the predictive time horizon, for weekends and weekdays.

Percent error was chosen as a measure of predictive accuracy for the different baseline approaches. It describes the relative magnitude and direction of the bias for each hour of the prediction and addresses the needs of the first two research questions this paper aims to address. The median percent error (MdPE) of the dataset is evaluated since it is less sensitive to the impact of outliers and presents a range of errors over the hours of the year. The MdPE computed is given in mathematical form in equation 1:

$$\mathrm{MdPE} = \mathrm{Median}_{\mathrm{k}} \frac{\left(E_{i,j} - \widehat{E}_{i,j}\right) x100}{E_{i,j}} \quad (1).$$

where i = 1,2,…2190 are the total hours within one season, j = 1,2,…120 are the number of meters and k = 0,1,2,…24 is the hour of day within a season. E is the actual energy consumption in the prediction period, $\widehat{E}$ is the predicted energy consumption. Table 2 provides the characterization of seasons that are applied in the analyses, based on the meteorological definitions of seasons provided in [23] (Trenberth, K. E. 1983).

The hourly MdPE value is represented by a heatmap, which is a two-dimensional graphical representation of data where the values are shown as colors. In the heatmap white represents zero MdPE, orange represents a positive MdPE (prediction is lower than actual consumption) and blue represents a negative MdPE (prediction is higher than actual consumption).

4.1. Data attrition method to assess baseline erosion due to frequency of implemented DF

To assess the degradation of predictive accuracy due to baseline erosion, we designed and tested a methodology that quantifies how the predictive accuracy changes as the baseline period training data is moved farther and farther away in time from the prediction day (to represent a hypothetical case where load shift is deployed on a series of consecutive days). This methodology is comprised of five steps:

i) Define a prediction period from available data: The 8 am-10 pm prediction window in the month of April was selected as the prediction period. The specific time window and the month of April was selected as it is an example of a spring month during which curtailment of solar generation is increasing ([9,24,25], and therefore of interest for load shift.

ii) Select appropriate baseline data: Prior to a prediction period on the first day in April, select appropriate number of baseline days for each baseline model. We ran three industry used baseline models: DM, TOWT7 and TOWT70. These models were a subset of the models described in section 2.1, representing two different approaches - averaging and regression.

iii) Generate hourly predictions: For the prediction period on the first day in April, generate hourly predictions for each model type.

iv) Move training data: Move baseline model training data back (initial test case has zero days removed from the baseline period, and subsequent test cases incrementally increase to removing 50 days) and for each day moved back repeat steps ii and iii. Fig. 1 represents

Table 2
Meteorological seasons.

Season	Months
Autumn	September, October, November
Spring	March, April, May
Summer	June, July, August
Winter	December, January, February

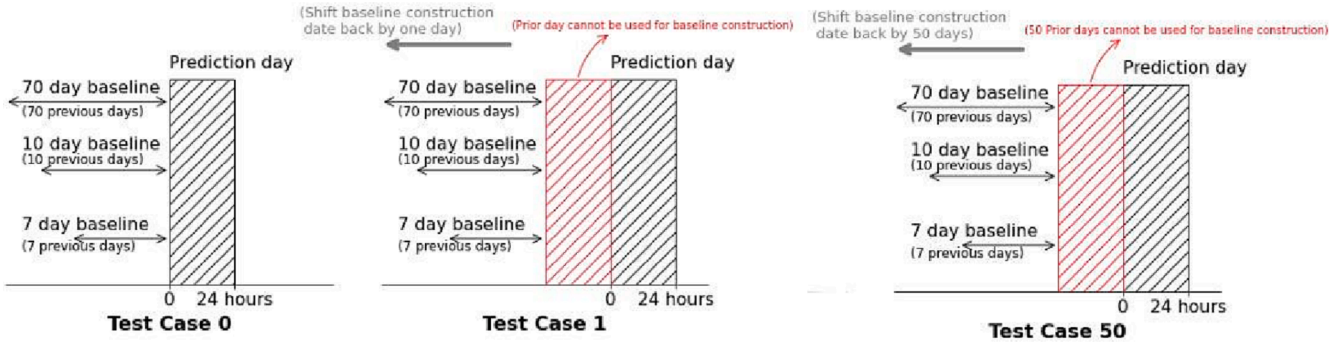


Fig. 1. Example test cases showing how the days before prediction day cannot be used for baseline model construction as baseline erodes.

examples of three different test cases revealing that as baseline erodes, the corresponding days before prediction day cannot be used for baseline model construction.

v) Assess predictive accuracy: In contrast to MdPE which was selected as an appropriate metric for assessing single-hour accuracy, the normalized mean bias error (NMBE) is more appropriate for assessing predictive accuracy across a multi-hour period (8:00am – 10:00 pm in this case). NMBE quantifies the mean bias error by dividing it by the mean of the actual load, giving the global difference between the actual and predicted load. Compute the NMBE performance metric for the predictions generated in the previous step for each test case across all trained models. The NMBE metric is also familiar with practitioners (ASHRAE Guideline 14 2014) and is independent of the timescale on which it is evaluated (Granderson 2016).

Equation (2) provides the mathematical form of the NMBE, where E_i is the actual load, $\widehat{E}_i$ the predicted load and $\overline{E}$ the mean. In the notation i = 1,2, 3,…N are the number of hours in the selected prediction window on each prediction day, j = 1…n number of meters in the dataset. The NMBE is computed for each test case.

$$NMBE = \left[\frac{\frac{1}{n}\Sigma_i^N (E_i - \widehat{E}_i)}{\overline{E}}\right] j \times 100 \tag{2}$$

vi) Assess degradation: Repeat steps ii, iii, iv and v for all days in April and compare the median and distribution of the NMBE error metric results across all meters and days as the baseline erosion period varies from zero to 50 days.

5. Results

This section provides results to address the research questions related to hourly predictive accuracy and baseline erosion.

5.1. Hourly predictive accuracy by season

The hourly MdPE in the prediction period is computed from the baseline model predictions, firstly for each season, and secondly for a month of April, a candidate month for potential LS applications [9,24,25]. The Fig. 2(a-f) shows the MdPE for each hour in the different seasons on weekdays. In addition, the figures in Appendix A summarize the 25th, 50th and 75th percentiles of the percent error for the models for each season and hour of the day and provide an insight into the performance of the models for the entire distribution. For the models and dataset used, the day matching (DM) model predicted all hours of day across all seasons with the smallest errors. Across all seasons of the year and all hours of the year, the largest median percent error for this model was −2.4% (hour 6 in autumn), and the smallest was −0.02% (hour 11 in spring). Across all models and hours on weekdays, summer was the most predictable season with the smallest errors and dispersion,

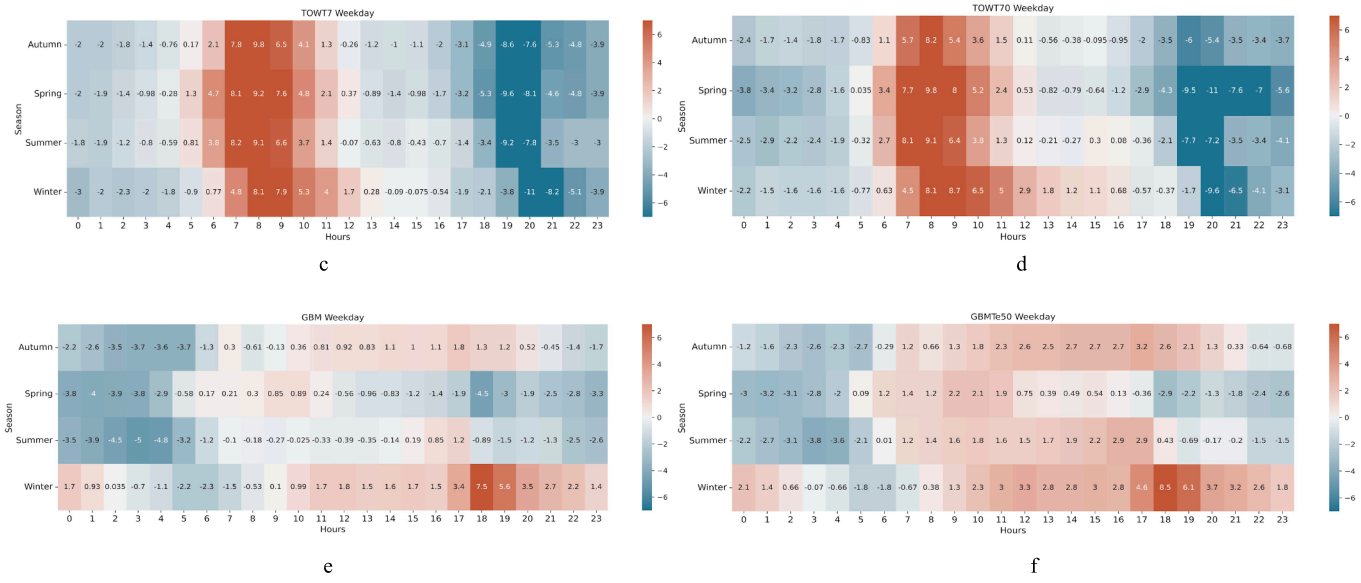


Fig. 2. Seasonal temporal variations of MdPE by model on weekdays. A. DM b. WM c. TOWT7 d. TOWT70 e. GBM f. GBMTe50.

followed by autumn. We also observe that a majority of hours in the spring season are over predicted across all models (indicated by more blue shading of the heat map charts).

Across all models on weekdays the afternoon hours between 12 and 16 tended to be most predictable, that is, predicted with smaller errors and dispersion. The evening hours 18–21 tended to be least predictable, with larger errors and dispersion. Across all models on weekdays, we observe a tendency to underpredict in the afternoon periods in the summer season, which is consistent with the findings in [19]. The regression models, TOWT7 and TOWT70 consistently under predict in the earlier hours of the day (7–10) and over predict in the later hours of the day (19–22). Both climate zones represented in the dataset in this study have a more distinct change in temperature during those two time periods (7–10, 19–22) [26,27], and these regression model residuals are autocorrelated, heteroscedastic and the regression parameters are correlated [21]. These model characteristics, coupled with the temperature characteristics of the dataset used could be why there is a distinct under and over prediction bias in those hours. The Fig. 3 (a-f) and Appendix B, summarize the hourly model MdPE in the different seasons for each hour of the day on weekends. In general, we observed lower MdPE for all models on weekends, as compared to weekdays, indicating that weekends tend to be more predictable.

5.2. Hourly predictive accuracy for an illustrative load shifting case

Fig. 4 presents a 95% confidence interval (CI) of the MdPE metric for the month of April for all models. While the MdPE provides a single estimate on the basis of the observed values of the statistic, the width of the CI indicates an interval estimate that specifies a range of values on either side of the MdPE within which the parameter can fall with a 95% level of confidence. We observe that the TOWT and TOWT70 have the widest CI as compared to other models, even though overall the CIs overlap considerably. For the April month, we observe similar behaviors as for the aggregated annual/seasonal results in Section 3.1, viz 1) the DM model has the least intraday variation and smallest errors, compared to other models, varying from −0.015 to −3.4 across the day. 2) The TOWT7 and TOWT70 have higher intraday variability and higher errors, while GBM and its variants tend to show lower variability and smaller errors. 3) Results also show that except for TOWT7 and TOWT70, all the models over predict (i.e., have negative MdPE values) in April for a majority of the time. When all models were compared to each other, the highest over prediction was seen in the 20th hour for the TOWT70 model with an MdPE of −12.5 and the highest under prediction was seen in the 8th hour with the TOWT7 model. Between the 11th and 16th hours of the day most models were close to the zero MdPE level but later in the day after the 16th hour their MdPE values showed more variability except for the DM. This is consistent with the results seen in the seasonal analysis for the season of Spring, where models over predicted 76% of the time.

Load shift may be quantified using 3 parameters - the load that is reduced ('shed') during one period of the day, the load that is increased ('take') in another period of the day, and the net change in load, i.e., the sum of the shed and the increase [29]. It is worth noting that the directionality and magnitude of model prediction bias can have differing impacts on each of the 3 parameters used to characterize and quantify the shift. Most of the models shown in Fig. 4 are consistently under-predicting building load. If applied to load shift this would overestimate the take portion, underestimate the shed portion, and consequently result in positive net change in load. In the case of the two TOWT-based models shown in Fig. 4 the inter-day change in directionality of bias may

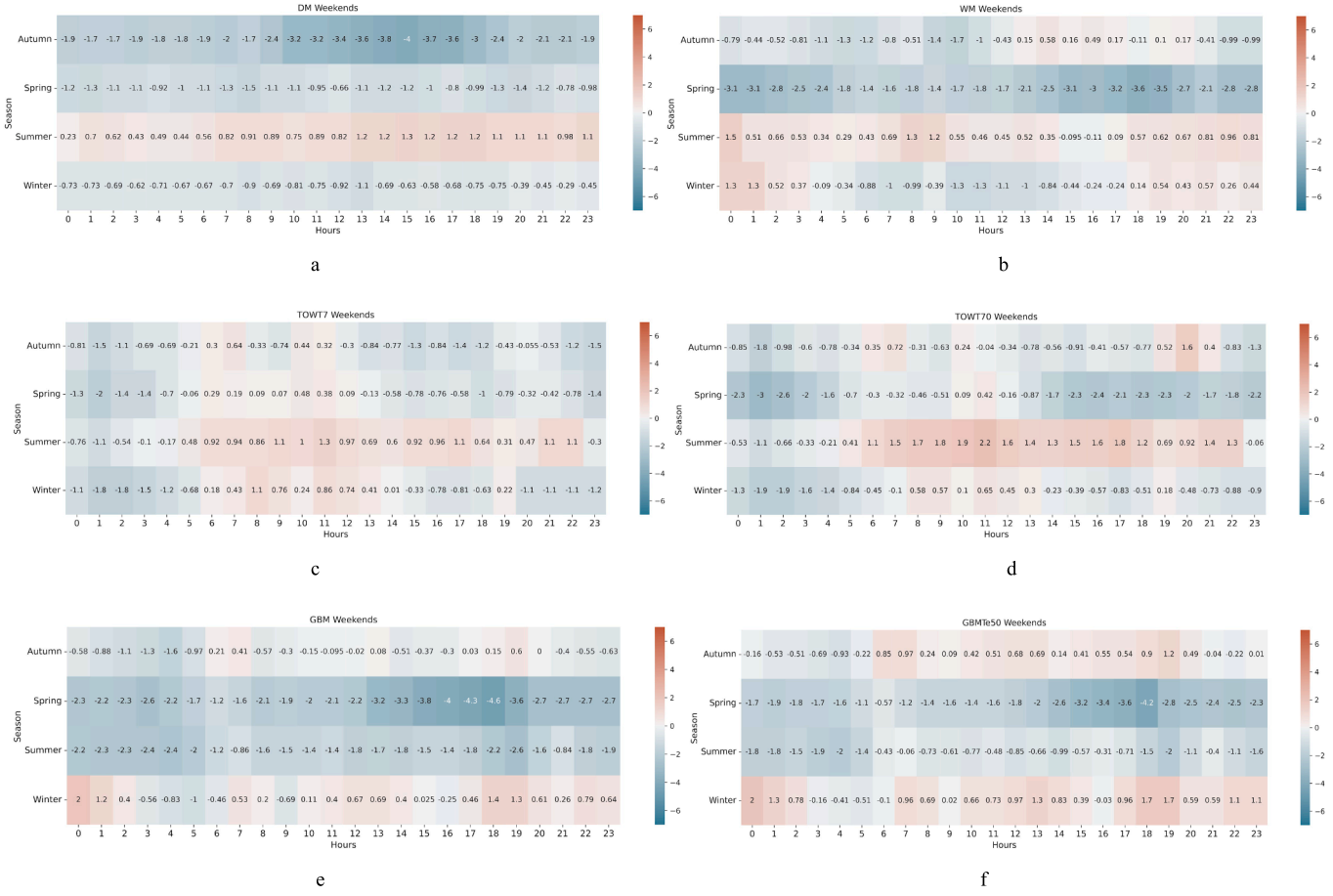


Fig. 3. Seasonal temporal variations of MdPE by model on weekends. a. DM b. WM c. TOWT7 d. TOWT70 e. GBM f. GBMTe50.

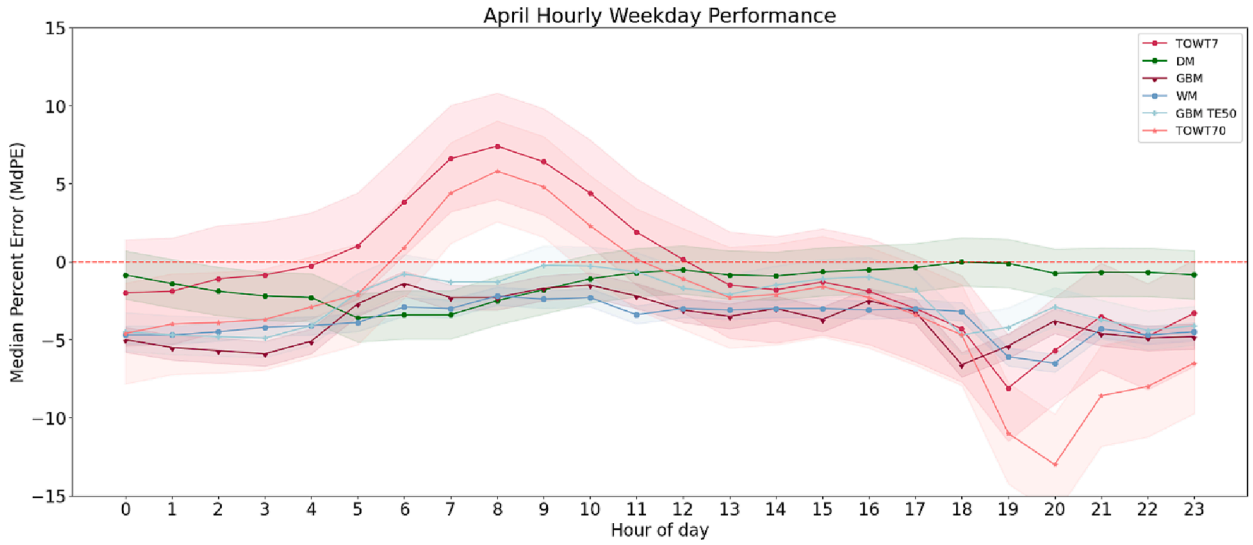


Fig. 4. April hourly weekday performance for models tested. Values above the red dotted line indicate under prediction by model, while values below indicate over prediction by a model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

actually reduce the error in the net calculation, as compared to a model with a consistent bias across the day (as seen with the matching and GBM models).

5.3. Baseline erosion and degradation of predictive accuracy

The box-and-whisker plots in Fig. 5a, 5c, 5e represents the NMBE across the full population of buildings in the test dataset for a time window of 8:00am to 10:00 pm, generally consistent with the hours of interest for daytime load shift. The median error is marked with a horizontal line within the interquartile range (IQR) box. The top of each 'whisker' represents the error for the 90th percentile in the population of test buildings and the bottom represents the 10th percentile. The top and bottom of each box represent the 75th and 25th percentiles, respectively. The Fig. 5b, 5d and 5f represents the IQR or dispersion plots for the three different models tested. The IQR is the difference between the

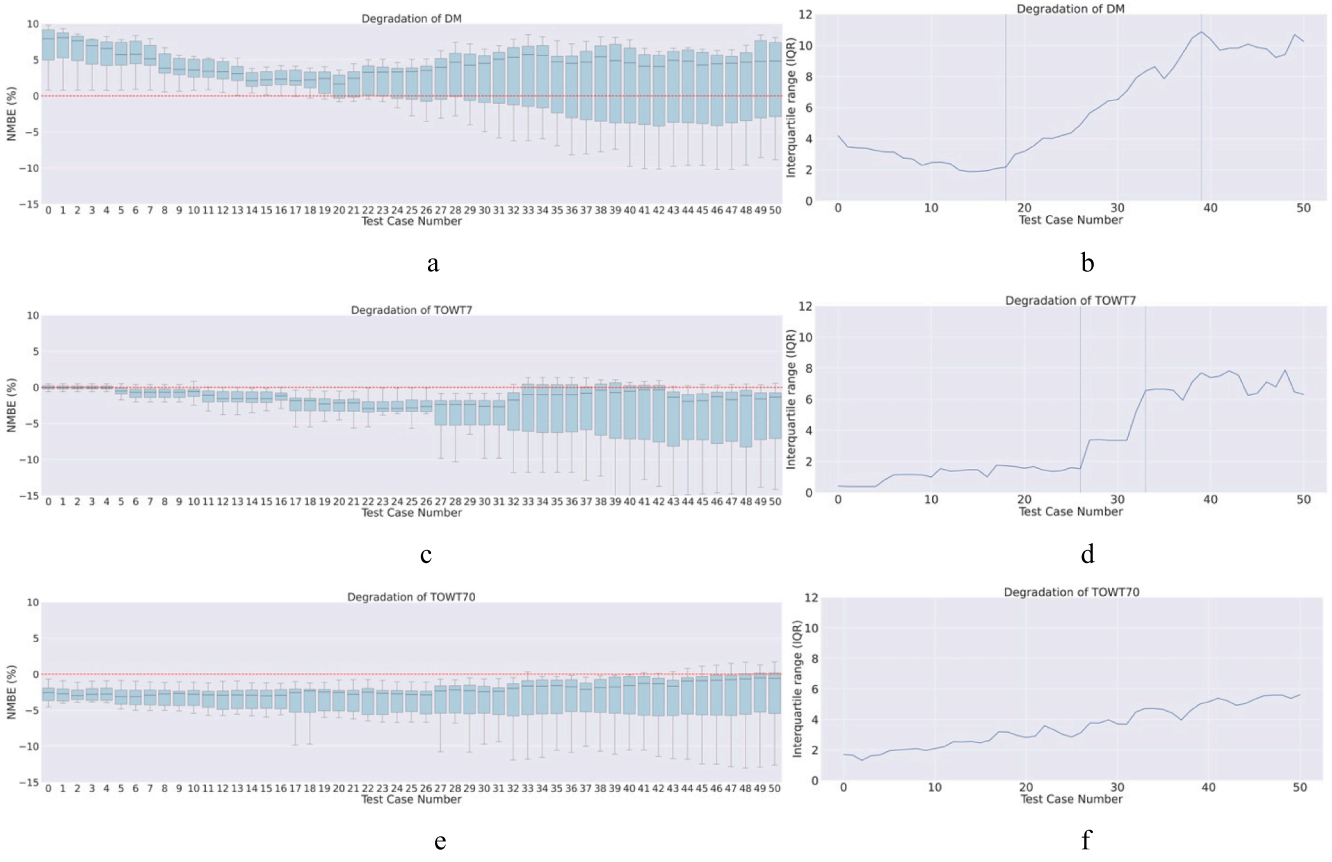


Fig. 5. Baseline erosion performance for each model for hours 8 am to 10 pm. Box-and-whisker plots on the left represent the NMBE across the full population of buildings, the right line chart represents the interquartile range (IQR) or dispersion plots for the three different models tested.

third quartile and first quartile and is a valuable tool for describing the spread of a given set of data.

We observe that for the DM model, the median NMBE has a wider range (8.1,2.2) compared to the other models. The IQR, while initially decreasing, shows an inflection point at the 18-day erosion point and increases with relatively constant slope until the 39-day erosion point before flattening again through the 50-day erosion point. For the TOWT7 model, the IQR increases steadily and then has an inflection point on the curve at the 25-day erosion point, thereafter increasing rapidly until the 34-day erosion point and flattening out. The TOWT70 model shows a gradual increase in median and IQR, but we do not observe any inflection point in the curve for any of the test cases as seen in the other two models. Across all models, we observe that as we progressively increase test cases (as baseline erodes), dispersion increases. In general, for the averaging algorithm with a 10 days baseline, we see higher dispersion than the regression approaches. Also, TOWT70, the model with most baseline days used in construction showed least dispersion as erosion increases, potentially because of its exposure to a wider range of conditions that were able to reflect those on the prediction days. The DM and TOWT7 may have been less likely to capture conditions that were sufficiently similar to the prediction day.

6. Discussion

The predictive accuracy of baseline energy or load models depends on many factors including model form, model training and prediction time horizons, granularity of the predicted quantity, and resolution of the training data. These factors interact differently, given the specific savings quantification application. For example, prior work has documented that to predict total energy use for a full year of consumption, as is used in many traditional efficiency savings applications, one year of training data drives higher accuracy than using 3 months of training data, and 15-minute interval training data drives higher accuracy than using monthly training data [16]. In the same vein, a day matching model with 10 days of baseline data may predict more accurately than certain regression models (Granderson et a. 2021), for annual peak load days, as in summer demand response applications. Each of these insights was derived from the development of predictive accuracy assessment methods tailored to specific use cases.

In this work we have presented and demonstrated new methods to assess predictive accuracy for applications that seek to quantify savings impact for specific hours of the day and different times of year. In the first example we demonstrate predictive accuracy for common industry baseline models for each of the 24 h in the day, for each season of the year. Hourly load predictions were generated for each day across a full year. A single month from the spring season, April, was also inspected. This analysis scenario aligns with a case in which total annual savings are of interest, as well as seasonal or monthly savings shapes, and desire to understand the extent to which load may have been shifted from one time of day to another through the measures implemented. In this case, the analysis revealed insights such as hours of the day (afternoons), periods of the week (weekends), and seasons (summer) that were predicted with more accuracy than other time periods. Across a population of commercial buildings, weekends are predominantly unoccupied hence these periods have a more weather-dependent load profile, which make them more predictable compared to weekdays when occupancy (a significant driver of load) can vary. In addition, the analysis showed tendencies toward overprediction (across most hours and models in spring) or underprediction (in the later hours of the day by some models in spring) of load, and how intra-shift changes in bias can impact quantification of net load shift. In this study we reported MdPE to provide a generalized comparison between algorithms, using data from several climate zones. For a specific load shift program application, it would be beneficial to conduct a similar assessment using building meter data from that specific region (and potentially limited to certain market segments if programs are sector-targeted). The magnitude of the median bias values observed are consistent with those seen in other related work such as Bode, J. et al. 2017, [19], EnerNOC 2011 (Bode et al. 2017, [19], EnerNOC [12,3]. It is noted that, while this paper focused on baseline prediction error at the level of an individual meter/building, the application of aggregated approaches across multiple meters would be expected to exhibit less bias, where those methods are applicable (e.g., for utility program evaluation)..

To understand baseline erosion, we analyzed a worst-case, or upper-bound degradation scenario in which there were up to 50 days separating the baseline period and the prediction day. In this scenario April is again used as the month of study, and predictive accuracy is assessed for a time window of hours 8–22, generally consistent with the hours of interest for daytime load shifting. As expected, it was observed that as the baseline period is pushed back further from the prediction day, the distribution of NMBE values for 120 m' baseline model predictions increases significantly. The regression model that used a 70-day baseline was more robust to baseline erosion, potentially because of its exposure to a wider range of conditions that were able to reflect those on the prediction days. Conversely, the 10-day matching baseline and the 7-day regression may have been less likely to capture conditions that were sufficiently similar to the prediction day. Unexpectedly, the bias of the day matching algorithm actually reduced over a two-week erosion period (both the median and IQR), before rapidly degrading. This behavior may be an artifact of the particular data set used, the particular model form, or a combination of the two. While the algorithms' predictive degradation varied, all performed relatively consistently across the initial 10–20 days, which is somewhat encouraging considering the assessment approach was based on a worst-case scenario of many consecutive days' erosion. Moreover, we observed that models with lower overall bias in hourly predictions were not necessarily those that were most robust to baseline erosion. The baseline models for load prediction in this work use historical electric load data and variables of the time of week and weather to make forecasts of load. The baseline erosion analysis revealed that the algorithms' predictive performance degraded considerably after the 20-day period and that baselines with an exposure to a wider range of conditions degraded less. Insights from a demand response program such as the number of event calls, timing and duration of event, extent of curtailment could inform future research into acceptable bounds for model prediction bias and for baseline degradation expectations.

7. Conclusion

Our work highlights the importance of looking at prediction algorithms differently for time-varying applications. We observe hourly and seasonal variation in bias, and existing predictive accuracy metrics do not sufficiently capture differences between approaches. We also provide the first published investigation of baseline erosion from repeated dispatch of building load shifting. Models that capture conditions similar to the prediction day perform better and we observed an inflection point in the curve for certain models that reveal days after which baseline erodes significantly.

The importance of quantifying time-resolved (hourly, monthly, and seasonal) changes in building energy use and demand is increasing as we recognize time-sensitive differences in the value of efficiency valuation, and as we use buildings to deliver dynamic load flexibility in support of the decarbonization goals. This work adds to the growing body of work in this area. Although summer peak-demand response programs have been delivered for years, other forms of solutions such as load shifting, and intentional delivery of efficiency measures that maximize grid and emissions benefits are newly emerging. As their associated market applications and implementation models are further defined, the methods demonstrated in this paper can be implemented to ensure that the most robust impact estimations are being used. That is, it will be possible to assess the predictive performance of a variety of load prediction algorithms in relation to the specific prediction horizons, hours of day and

times of year applicable to different DF strategies. Such program-aligned analyses are a focus of future work. Future work may also consider extension of these approaches to the quantification of greenhouse gas emissions reductions.

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.

Data availability

Data will be made available on request.

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. The authors would like to thank Sarah Zaleski and Marc LaFrance from the U.S. Department of Energy Building Technologies Office for their support of this work.

Appendix A. Distribution of percent error on weekday for each model by season

Hourly distribution of percent error by model and season for weekdays. a. DM b. WM c. TOWT7 d. TOWT70 e. GBM f. GBMTe50.

Appendix B. Distribution of percent error on weekend for each model by season

Hourly distribution of percent error by model and season for weekends. g. DM h. WM i. TOWT7 j. TOWT70 k. GBM l. GBMTe50.

References

[1] American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE). (2014). ASHRAE Guideline 14-2014 for Measurement of Energy and Demand Savings, American Society of Heating, Refrigeration and Air Conditioning Engineers, Atlanta, GA.
[2] American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE). (2021). ASHRAE standard 169-2021, Climate Data for Building Design Standards. Atlanta, Ga.
[3] EnerNOC Inc The demand response baseline https://bit.ly/3zQ7ldl 2011 Accessed 28 October 2022.
[4] K. Coughlin, M.A. Piette, C. Goldman, S. Kiliccote, Statistical analysis of baseline load models for non-residential buildings, Energy and Buildings 41 (4) (2009) 374–381, https://doi.org/10.1016/j.enbuild.2008.11.002.
[5] M. Neukomm, V. Nubbe, R. Fares, Grid-interactive efficient buildings technical report series: Overview of research challenges and gaps, United states (2019), https://doi.org/10.2172/1580214.
[6] Bode, J., & Ciccone, A. (2017). California ISO Baseline Accuracy Assessment. https://bit.ly/3Ufd5FK Accessed 28 October 2022.
[7] Franconi, E., Gee, M., Goldberg, M., Granderson, J., Guiterman, T., Li, M., and Smith, B.A. (2017). Advanced M&V Status and Prospects: "M&V 2.0" Methods, Tools, and Applications, Lawrence Berkeley National Laboratory, February 2017. LBNL report number LBNL-1007125.
[8] U.S. Energy Information Administration (EIA). Frequently Asked Questions (FAQs). (n.d). https://bit.ly/3UgYvNW . Accessed 1 Nov. 2022.
[9] U.S. Energy Information Administration (EIA). California's curtailments of solar electricity generation continue to increase. (2021). https://bit.ly/3Eafxrd . Accessed 1 November 2022.
[10] Irena. (2020). Global renewables outlook: Energy transformation 2050.https://bit.ly/3NMJ4dW . Accessed 31 October 2022.
[11] R. Wiser M. Bolinger B. Hoen D. Millstein J. Rand G. Barbose B. Paulos Land-based wind market report 2022 edition. 2022 Lawrence Berkeley National Lab (LBNL), Berkeley, CA (United States).
[12] K. Inc PJM Empirical Analysis of Demand Response Baseline Methods https://bit.ly/3hlseqi 2011 Accessed 31 October 2022.
[13] J. Cowan, International performance measurement and verification protocol: Concepts and Options for Determining Energy and Water Savings-Vol, I. International Performance Measurement & Verification Protocol 1 (2002).
[14] J. Granderson, P.N. Price, D. Jump, N. Addy, M.D. Sohn, Automated measurement and verification: Performance of public domain whole-building electric baseline models, Applied Energy 144 (2015) 106–113.
[15] Granderson, J., Touzani, S., Custodio, C., Sohn, M. D., Jump, D., & Fernandes, S. (2015b). Accuracy of automated measurement and verification (M&V) methods. https://bit.ly/3hor1yo.Accessed 31 October 2022.
[16] J. Granderson, S. Touzani, C. Custodio, M.D. Sohn, D. Jump, S. Fernandes, Accuracy of automated measurement and verification (M&V) techniques for energy savings in commercial buildings, Applied Energy 173 (April) (2016) 296–308, https://doi.org/10.1016/j.apenergy.2016.04.049.
[17] J. Granderson, S. Fernandes, S. Touzani, C.C. Lee, E. Crowe, M. Sheridan, Spatio-temporal impacts of a utility's efficiency portfolio on the distribution grid, Energy 212 (2020), 118669, https://doi.org/10.1016/j.energy.2020.118669.
[18] Schiller, S. R., Schwartz, L. C., & Murphy, S. (2020). Performance Assessments of Demand Flexibility from Grid-Interactive Efficient Buildings: Issues and Considerations (No. DOE/EE-2087). Lawrence Berkeley National Lab(LBNL), Berkeley, CA (United States). https://doi.org/10.2172/1644287.
[19] J. Granderson, M. Sharma, E. Crowe, D. Jump, S. Fernandes, S. Touzani, D. Johnson, Assessment of Model-Based peak electric consumption prediction for commercial buildings, Energy and Buildings 245 (2021), https://doi.org/10.1016/j.enbuild.2021.111031.
[21] J.L. Mathieu, P.N. Price, S. Kiliccote, M.A. Piette, Quantifying changes in building electricity use, with application to demand response, IEEE Transactions on Smart Grid 2 (3) (2011) 507–518, https://doi.org/10.1109/TSG.2011.2145010.
[22] S. Touzani, J. Granderson, S. Fernandes, Gradient boosting machine for modeling the energy consumption of commercial buildings, Energy and Buildings 158 (2018) 1533–1543, https://doi.org/10.1016/j.enbuild.2017.11.039.
[23] K.E. Trenberth, What are the seasons? Bulletin of the American Meteorological Society 64 (11) (1983) 1276–1282.
[24] Koolbeck,M., Lezaks,J., Shaver,L. (2020). Market potential for saving energy and carbon emissions with load shifting measures. https://bit.ly/3EboGi6 . Accessed on October 27, 2022.
[25] California Public Utilities Commission (CPUC) Final report of the CPUC working group on load shift https://bit.ly/3UTTsTx 2019 Accessed on October 27, 2022.
[26] W. Inc Average hourly temperature in California http://bit.ly/3UY7dRq 2022 Accessed on November 22, 2022.
[27] W. Inc Average hourly temperature in Washington DC http://bit.ly/3VnlSFs 2022 Accessed on November 22, 2022.
[29] J. Liu, R. Yin, L. Yu, M.A. Piette, M. Pritoni, A. Casillas, J. Xie, T. Hong, M. Neukomm, P. Schwartz, Defining and applying an electricity demand flexibility benchmarking metrics framework for grid-interactive efficient commercial buildings, Advances in Applied Energy 8 (2022) 100107.
[30] N. Mims Frick, L. Schwartz, Time-sensitive value of efficiency: Use cases in electricity sector planning and programs, Berkeley, CA, USA, LBNL, 2019.
[31] S. Murphy, J. Deason, A. Satchwell, Timed to save: the added value of accounting for hourly incidence of electricity savings from residential space-conditioning measures, Energy Efficiency 14 (8) (2021) 1–13.
[32] A. Poulin, M.A. Leduc, M. Fournier, Statistical Analysis of Baseline Load Models for Residential Buildings in the Context of Winter Demand Response, Energies 15 (12) (2022) 4441.
[33] L. Hatton, P. Charpentier, E. Matzner-Lober, Statistical estimation of the residential baseline, IEEE Transactions on Power Systems 31 (3) (2015) 1752–1759.