Applied Energy 173 (2016) 296–308

Contents lists available at ScienceDirect

Applied Energy

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

Accuracy of automated measurement and verification (M&V) techniques for energy savings in commercial buildings

Jessica Granderson [a,*], Samir Touzani [a], Claudine Custodio [a], Michael D. Sohn [a], David Jump [b], Samuel Fernandes [a]

[a] Lawrence Berkeley National Laboratory, 1 Cyclotron Rd., Berkeley, CA 94720, USA
[b] Quantum Energy Services and Technologies, Inc., 2001 Addison St., Suite 300, Berkeley, CA 94704, USA

HIGHLIGHTS

- A testing procedure and metrics to asses the performance of whole-building M&V methods is presented.
- The accuracy of ten baseline models is evaluated on measured data from 537 commercial buildings.
- The impact of reducing the training period from 12-months to shorter time horizon is examined.

ARTICLE INFO

Article history:
Received 1 October 2015
Received in revised form 28 March 2016
Accepted 10 April 2016
Available online 16 April 2016

Keywords:
Baseline model
Measurement and verification
Whole-building energy
Predictive performance accuracy
Building energy analysis
M&V 2.0

ABSTRACT

Trustworthy savings calculations are critical to convincing investors in energy efficiency projects of the benefit and cost-effectiveness of such investments and their ability to replace or defer supply-side capital investments. However, today's methods for measurement and verification (M&V) of energy savings constitute a significant portion of the total costs of efficiency projects. They also require time-consuming manual data acquisition and often do not deliver results until years after the program period has ended. The rising availability of "smart" meters, combined with new analytical approaches to quantifying savings, has opened the door to conducting M&V more quickly and at lower cost, with comparable or improved accuracy. These meter- and software-based approaches, increasingly referred to as "M&V 2.0", are the subject of surging industry interest, particularly in the context of utility energy efficiency programs. Program administrators, evaluators, and regulators are asking how M&V 2.0 compares with more traditional methods, how proprietary software can be transparently performance tested, how these techniques can be integrated into the next generation of whole-building focused efficiency programs.

This paper expands recent analyses of public-domain whole-building M&V methods, focusing on more novel M&V 2.0 modeling approaches that are used in commercial technologies, as well as approaches that are documented in the literature, and/or developed by the academic building research community. We present a testing procedure and metrics to assess the performance of whole-building M&V methods. We then illustrate the test procedure by evaluating the accuracy of ten baseline energy use models, against measured data from a large dataset of 537 buildings. The results of this study show that the already available advanced interval data baseline models hold great promise for scaling the adoption of building measured savings calculations using Advanced Metering Infrastructure (AMI) data. Median coefficient of variation of the root mean squared error (CV(RMSE)) was less than 25% for every model tested when twelve months of training data were used. With even six months of training data, median CV(RMSE) for daily energy total was under 25% for all models tested. These findings can be used to build confidence in model robustness, and the readiness of these approaches for industry uptake and adoption.

1. Introduction

In 2003, 18% of the US total energy consumption was attributed to commercial buildings, amounting to approximately 17 quadrillion British thermal units [1]. To reduce the environmental and cost impacts associated with the buildings sector, utility demand

* Corresponding author at: 1 Cyclotron Rd., MS 90-3111, Berkeley, CA 94720, USA.
E-mail address: JGranderson@lbl.gov (J. Granderson).

http://dx.doi.org/10.1016/j.apenergy.2016.04.049
0306-2619/

side management programs were established. These programs have grown over time, representing a $7.2B investment as of 2012 [2]. Quantification of the savings that these programs achieve is a critical component in evaluating their impact. Savings quantification is also used in the energy services company (ESCO) industry, and in projects self-financed by building owners. Similar in magnitude to the utility program industry, the annual revenues of the US ESCOs industry were evaluated at around $7 billion with roughly 75% associated with energy efficiency projects [3].

In these industries, "measurement and verification" is the process of estimating savings relative to a baseline period, and is therefore critical to establishing the value of efficiency to building owners, utility rate payers, and service providers. However, M&V can be quite costly and time consuming, with questions remaining as to the accuracy of the estimated savings. Depending on the M&V methods employed and whether third party evaluation is included, M&V costs can range from 1% to 5% of project portfolio costs [4]. Today, the growing availability of data from smart meters and devices, combined with advanced data analytics offers the potential to streamline the M&V process through increased levels of automation, while maintaining or improving the accuracy of the result.

Many of the technologies included in the efficiency strategies, such as EMIS (Energy Management and Information Systems) include building energy baseline modeling functionality that can be used to automatically quantify savings [5]. Automated quantification of savings is currently available in a range of energy management tools, including onsite or software-as-a-service software offerings that track monthly or interval energy consumption for individual sites or portfolios of buildings. A recent study by Portland Energy Conservation Inc. (PECI) for the Northwest Energy Efficiency Association documented commercial energy management tools with functionality for M&V applications [6]. Currently, M&V 2.0 and technologies such as EMIS, are receiving an unseen level of attention in the industry, due to their promise to reduce program time and costs, and unlock untapped savings through whole-building focused programs. For example, recent papers of two leading US efficiency organizations, the American Council for an Energy Efficient Economy, and Northeast Energy Efficiency Partnerships, have highlighted both the promise of, and need for a better understanding of the performance of, and practical uses of these emerging approaches to savings estimation and evaluation [7,8].

The baseline models used in M&V 2.0 are empirical models that relate energy usage to parameters such as outdoor air temperature, humidity or building operating schedule. These models, which are developed using the pre-retrofit data, are used to estimate the energy use in the post-retrofit period. The difference between the estimated and the metered energy consumption is taken as the 'avoided energy use' or energy savings. Traditionally, monthly utility bills data were used to build the baseline models, however, the increasingly availability of hourly and 15-min interval meter data has enabled new models with the potential for more accurate M&V.

Several methods of baseline modeling that use interval meter data have been recently introduced in the literature. The day–time–temperature regression model described in [9] include time of the day, day of the week, and two temperature variables to allow different heating and cooling slopes, it also can include humidity and holidays as variables. This model is fit with ordinary least square regression. The Time-of-Week-and-Temperature model, which is described in [10,11] and used in this study, is a regression model that includes time of week, and a piecewise-continuous temperature response with several change points. A weighted version of this model is proposed in [12]. In [13] the authors utilized the Gaussian mixture models for modeling the energy of commercial buildings. In the framework of Bayesian statistics a Gaussian process model was introduced in [14] for estimating the energy use of buildings. See [15] for a review of other baseline models and modeling approach for prediction of building energy consumption.

Although these emerging analytical methods, and the EMIS technologies in which they are automated hold great promise in reducing the cost and time required for M&V in the commercial buildings sector [7,8], several questions relating to their use remain to be answered, for example:

- What metrics should be used to quantify the performance of these tools?
- How accurate are automated baseline models that utilize interval meter data?
- How can the performance of proprietary tools that automate gross savings calculations be evaluated?
- How can one tool or model be compared to another?

While resources such as the IPMVP [16] and ASHRAE Guideline 14 [17], establish procedural and quantitative requirements for baseline model construction, goodness-of-fit to data during the model training period, and rules of thumb for model application given different expected depths of savings, they do not provide a general means of assessing model performance during a *prediction* period. The testing procedure presented in this work extends the principles in these existing industry resources to quantify model predictive accuracy beyond the training period. As noted in Section 4, the evaluation of goodness-of-fit, which is measured on the training period, is related to, but not an explicit indication of the accuracy of actual model predictive performance.

In this paper, we expand prior work to answer these questions. Specifically, we present five key outcomes: (1) a test procedure used to evaluate the accuracy of automated building energy baseline models that are used in avoided energy use calculations to determine what the building would have consumed had no efficiency measure been installed; (2) the application of that test procedure to evaluate ten novel interval data-based models, using metered data from hundreds of geographically diverse buildings; (3) two stakeholder consensus-based performance metrics; (4) interpretation of model performance and discussion of implications for the M&V industry; and (5) conclusions and directions for future work. While prior work [10,18] focused on a limited number of both monthly and interval data models that are published in the literature, this study analyzes the performance of an expanded set of models from commercial service and tools providers, and the research community. In addition, prior work did not include efforts to engage the stakeholder community to identify consensus-based performance metrics, and used more geographically limited test data sets. In presenting consensus-based performance metrics, an evaluation of models from proprietary commercial tools, and testing a set of novel interval baseline models, this work addresses key questions currently being asked across the US utility program community – the answers to which are critical in leveraging new technology to advance the state of practice in the efficiency programs industry. Combined with prior work, this more recent study provides strong evidence for the promise of these emerging M&V methods to streamline the M&V process, and to facilitate increased adoption in industry efficiency applications.

The analyses that are presented in this paper represent a 'floor' for predictive accuracy, using fully automated approaches. Data was provided to the models, which automatically fit their parameters, and model-predictions were compared to actual meter data. No attempt was made to implement non-routine adjustments to improve model predictions. Therefore, the accuracy results that

are presented represent the most conservative view into performance, which could be improved with the oversight of an engineer. The vision motivating this work can be understood in general: that by using large test data sets, predictive accuracy can be verified for large portions of building populations. This performance validation can provide the quantitative evidence and confidence to begin leveraging automation to scale: (a) the adoption of measured pre/post M&V approaches, and (b) the number of buildings for which M&V can be conducted with decreased time and cost. Energy efficiency savings that is verified within specific known error bounds may be more interesting as a commodity to some potential buyers.

2. Methodology

2.1. Overview

The evaluation of model predictive accuracy that is presented in this paper is based on the refinement of a 4-step testing procedure discussed in [10], and is depicted in Fig. 1. The test dataset comprises interval meter data and an independent variable data, which is outside air temperature, for several hundreds of buildings. These buildings are "untreated" in terms of efficiency interventions. That is, they are not known to have implemented major efficiency measures.

The data for each building is divided into hypothetical training periods and prediction periods, and meter data from the prediction period is "hidden" from the model. The trained model is used to forecast the load throughout the prediction period, and predictions are then compared to the actual meter data that had been hidden. Fig. 2 shows an example of actual, and model-predicted data for a 12-month training period and a 12-month prediction period (in this example the prediction was performed by the Time-of-Week-and-Temperature model). Performance metrics that quantify the difference between the model prediction and the actual load are calculated and used to characterize accuracy. This test procedure is documented in further detail in previous publications [10,18–20]; in these publications, the similarities to the ASHRAE 'shootouts' of the mid and late 1990s [21,22] are noted, as well as key differences that represent an evolution of the overall body of work that considers the performance of energy baseline models.

An important feature of this test procedure is that it can be used to assess the predictive accuracy of a model, objectively, without needing to know the specific algorithm or the underlying form of the model. Therefore, proprietary tools can be evaluated while protecting the developer's commercial intellectual property. In addition, it provides a general approach to evaluate the errors in calculated energy savings, according to diverse pre- and post-measure time horizons, and large test sets of building energy data.

2.2. Test data

The test dataset that was compiled for this analysis comprised whole-building data that represented a dataset of convenience, as opposed to one driven by an ideal experimental design. This is due to the well-known challenges associated with obtaining access to customer utility data. Ideally, the buildings would be uniformly distributed across all climate zones, or would have reflected a sampling strategy based on the intended use of the results; however it was not possible to obtain that level of diversity for this study. The data that were acquired were skewed to buildings from California, and Washington, DC, with much less representation from other regions. Hence, the test dataset for the analyses presented in this paper comprised 537 commercial buildings from multiple ASHRAE climate zones [23], and is characterized in Table 1. For each building, 15-min whole-building electricity data was paired with outside air temperature that was determined from the building's zip code. Buildings in ASHRAE Climate Zone 3 were from Northern and Central California and those from Climate Zone 4 were from the Northwest and Mid-Atlantic regions. Fig. 3 shows the ASHRAE

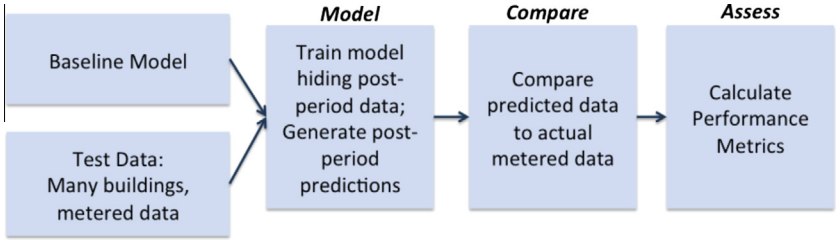

Fig. 1. Schematic of the general methodology used to evaluate the performance of automated M&V methods.

Fig. 2. Actual and model-predicted energy data, overlaid with outside air temperature, for a 12-month training period and 12-month prediction period.

Table 1
Summary of Climate Zones of buildings used to test model performance.

ASHRAE Climate Zone	1 (Very Hot)	2 (Hot)	3 (Warm)	4 (Mixed)	5 (Cool)	6 (Cold)	7 (Very Cold)
Building count	1	15	277	237	5	1	1

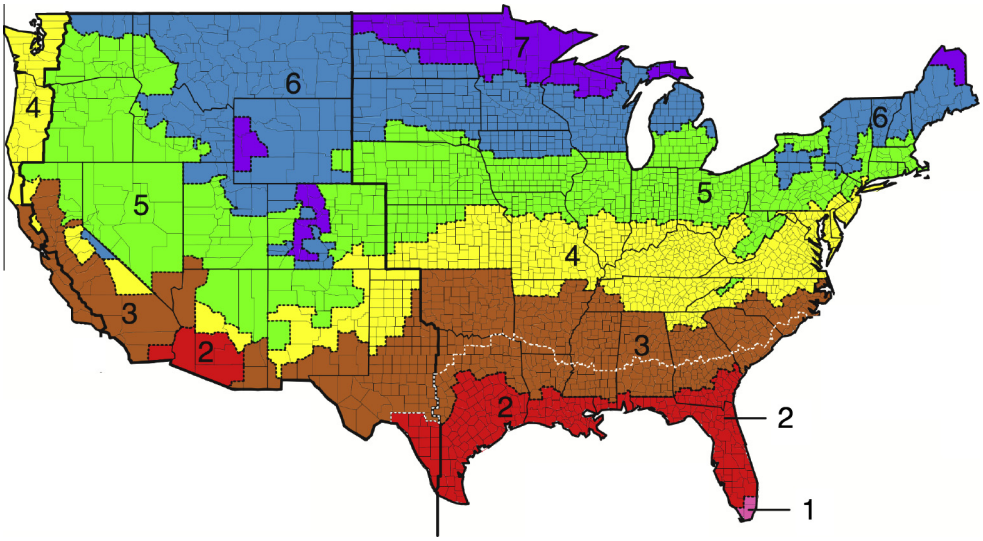

Fig. 3. US Map with ASHRAE-IECC Climate Zones [23].

Climate Zones overlaid on a map of the Unites States. The analyses presented in this work are constrained to data and models of whole building electric metering because it is the type of interval data most readily available in today's buildings.

2.3. Description of models tested

Ten baseline models were evaluated in this study, comprising a cross-section of approaches used in commercial EMIS technologies, as well as approaches that are documented in the literature, and/or developed by the academic building research community. The models that were selected are novel in that they move beyond simple change point and regression models, however they are not yet widely used in practice. The models are described below. While the models may be able to accommodate additional independent variables, outside air temperature, date, and time were the only variables for which it was possible to build a large dataset comprising hundreds of buildings from diverse climates. Some of these models are further explained in Appendix A.

M1. *Combination principle component analysis and bin modeling,* developed by Buildings Alive Pty. Ltd., of Sydney Australia.
M2. *Combination Random Forest* [24], *Extremely Randomized Trees* [25] *and Mean Week,* developed by Paul Raftery and Tyler Hoyt at the Center for the Built Environment, University of California, Berkeley.
M3. *Advanced regression including a term for drift,* developed by Gridium Inc.
M4. *Mean Week* – predictions depend on day and time only. For example, the prediction for Tuesday at 3 PM is the average of all of the data for Tuesdays at 3 PM. Therefore, there is a different load profile for each day of the week, but not, for example, for each week in a month or each month in the year. This is a simplistic 'naïve' model that was intentionally included for comparative purposes.
M5. *Time-of-week-and-temperature* [11]: the predicted load is a sum of two terms: (1) a "time of week effect" that allows each time of the week to have a different predicted load from the others, and (2) a piecewise-continuous effect of temperature. The temperature effect is estimated separately for periods of the day with high and low load, to capture different temperature slopes for occupied and unoccupied building modes.
M6. *Weighted time-of-week-and-temperature* [12]: the *time-of-week and-temperature* model with the addition of a weighting factor to give more statistical weight to days that are nearby to the day being predicted. This is achieved by fitting the regression model using weights that fall off as a function of time in both directions from a central day.
M7. *Ensemble approach combining nearest neighbors* [26] *and a generalized linear model* [27], developed by Lucid Design Group.
M8. *Combination Multivariate Adaptive Regression Splines* (*MARS*) [28] *and other advanced regression.*
M9. *Combination bin modeling and other advanced regression,* developed by Performance Systems Development of New York, LLC.
M10. *Nearest neighbor advanced regression.*

2.4. Performance metrics

There are many metrics that can be used to quantify the accuracy of model predictions. Different metrics provide different insights into aspects of performance. To identify those most relevant and useful in understanding model performance for M&V of energy savings, a group of approximately twenty industry representatives from the evaluation, implementation, and utility program management community were consulted. These stakeholders were asked to select from several candidates such as coefficient of determination, root mean squared error and other goodness-of-fit metrics. Across this group of experts, the two most meaningful for M&V applications were found to be the normalized

mean bias error (NMBE) and the coefficient of variation of the root mean squared error (CV(RMSE)). These two metrics provide complementary views of model performance for M&V applications. They also provide a means to assess relative model-to-model comparisons across several buildings simultaneously.

The NMBE is the mean of the error in the predictions divided by the mean of the actual energy use. In other words, it gives a sense of the total difference between model predicted energy uses, and actual metered energy use, with intuitive implications for the accuracy of avoided energy use calculations. If the value of NMBE is positive, it means that the prediction of the total energy used during the entire prediction period is lower than the measured value. A negative NMBE means that the prediction is higher. The NMBE is defined in the following equation, 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.

$$\mathrm{NMBE} = \frac{\frac{1}{n}\sum_i^n (y_i - \hat{y}_i)}{\bar{y}} \times 100 \quad (1)$$

The value of NMBE is independent of the timescale on which it is evaluated, which means that the value of the metric will be the same if the timescale is 15-min, hourly or daily.

The CV(RMSE) is the root mean square error divided by the mean of the measured values, which provide a quantification of the typical size of the error relative to the mean of the observations. This metric also gives an indication of the model's ability to predict the overall load shape that is reflected in the data. CV (RMSE) is also familiar to practitioners, and is prominent in resources such as ASHRAE Guideline 14. The CV(RMSE) is defined by the Equations 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.

$$\mathrm{CV(RMSE)} = \frac{\sqrt{\frac{1}{n}\sum_i^n (y_i - \hat{y}_i)^2}}{\bar{y}} \times 100 \quad (2)$$

In contrast to the NMBE, CV(RMSE) quantifies the predictive accuracy at the timescale of the data and prediction, in other words, if the predictions and measured data apply to 15-min then this metric summarizes the accuracy in 15-min predictions. In this study, in addition to the 15-min CV(RMSE) metric, the daily CV (RMSE) is also presented. The total predicted daily energy use across the prediction period is calculated using the results of the predictions based on the 15-min interval data.

2.5. Time horizons

In keeping with the current standard practice and guidelines for whole-building avoided energy use calculations [17], the analyses in this study are grounded in a 12-month 'post' or model prediction period. This prediction period corresponds to the last 12 month of the available data for each studied building. We assess the degradation in prediction accuracy when 'pre' or model training period is reduced from 12-months to shorter time horizons. Specifically, results are presented for 12-month, 9-month, 6-month and 3-month training periods. These training periods correspond to the periods that immediately precede the prediction period. Note that not all buildings from the test dataset had a full 24 months of electricity and outside air temperature data. Therefore, the models were tested on different numbers of buildings for each training period; for the 12-month, 9-month, 6-month and 3-month training periods the number of buildings were 441, 470, 530 and 537 respectively.

3. Results

Some buildings are predictable, and others are not; therefore, to understand the predictive accuracy of the models, and their promise for streamlining M&V, it is necessary to test them across *many* buildings. Moreover, simply reporting the mean or median does not give a full picture of the fraction of buildings in the population for which accuracy is exceptionally high or low; therefore the results present distributions, i.e., percentiles, of the performance metrics over the full population of buildings in the data set.

Most models were unable to generate predictions for at least some of the buildings in the data set – summarized in Table 2, failure rates ranged from roughly zero to ten percent depending on the training period and particular model in question. In the table, the total number of failures is shown first, with the percentage of failures (failed buildings divided by total buildings), is shown in parentheses. These aspects of performance are likely due to differences in the underlying form of the models, how they were coded to run automatically in batch mode, their treatment of outliers in the training data, and the different mathematical approaches that they each use.

3.1. Normalized mean bias error

Normalized mean bias error across the full population of buildings in the test dataset is shown for each model, in Fig. 4. In these box-and-whisker plots, the mean error is shown with a white circle; for some models, the mean error is literally off of the chart, and not plotted. 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; note that for some models, these two percentiles are also off of the chart, and thus not displayed. The top and bottom of each box represent the 75th and 25th percentiles, respectively, and the horizontal line in each box marks the median, or 50th percentile. The number of buildings in the test dataset by training period is shown in the title at the top of each plot.

While Fig. 4 shows percentiles of errors across the full population of buildings and training periods that were analyzed, Table 3 summarizes just the 25th, 50th (median) and the 75th percentiles error as the training period is reduced from twelve, to nine, to six, to three months. This provides insight into the general degradation in performance that is seen as the model training period is reduced, while the prediction period is held fixed at twelve months.

The results displayed in Fig. 4 and Table 3 show that for the majority of cases there was a tendency of a bias toward over-predicting the energy use (NMBE negative). However, this may be a result of actual decreases in building energy use over time, as opposed to a characteristic of the models. Further research is needed to explore this premise. In addition, when the training period was shortened from twelve months to nine and to six the

Table 2
Number of failures for each model, for a 12-month prediction period and 12-month, 9-month, 6-month, and 3-month training periods.

Model	12 months	9 months	6 months	3 months
#Buildings	441	470	530	537
M1	0 (0%)	0 (0%)	3 (0.57%)	4 (0.75%)
M2	26 (5.90%)	24 (5.11%)	34 (6.42%)	34 (6.33%)
M3	7 (1.59%)	15 (3.19%)	16 (3.02%)	13 (2.42%)
M4	0 (0%)	0 (0%)	0 (0%)	0 (0%)
M5	0 (0%)	0 (0%)	0 (0%)	0 (0%)
M6	0 (0%)	0 (0%)	0 (0%)	0 (0%)
M7	24 (5.44%)	37 (7.87%)	56 (10.57%)	38 (7.08%)
M8	8 (1.81%)	6 (1.28%)	18 (3.40%)	65 (12.10%)
M9	20 (4.54%)	4 (0.85%)	4 (0.75%)	4 (0.75%)
M10	0 (0%)	0 (0%)	0 (0%)	2 (0.37%)

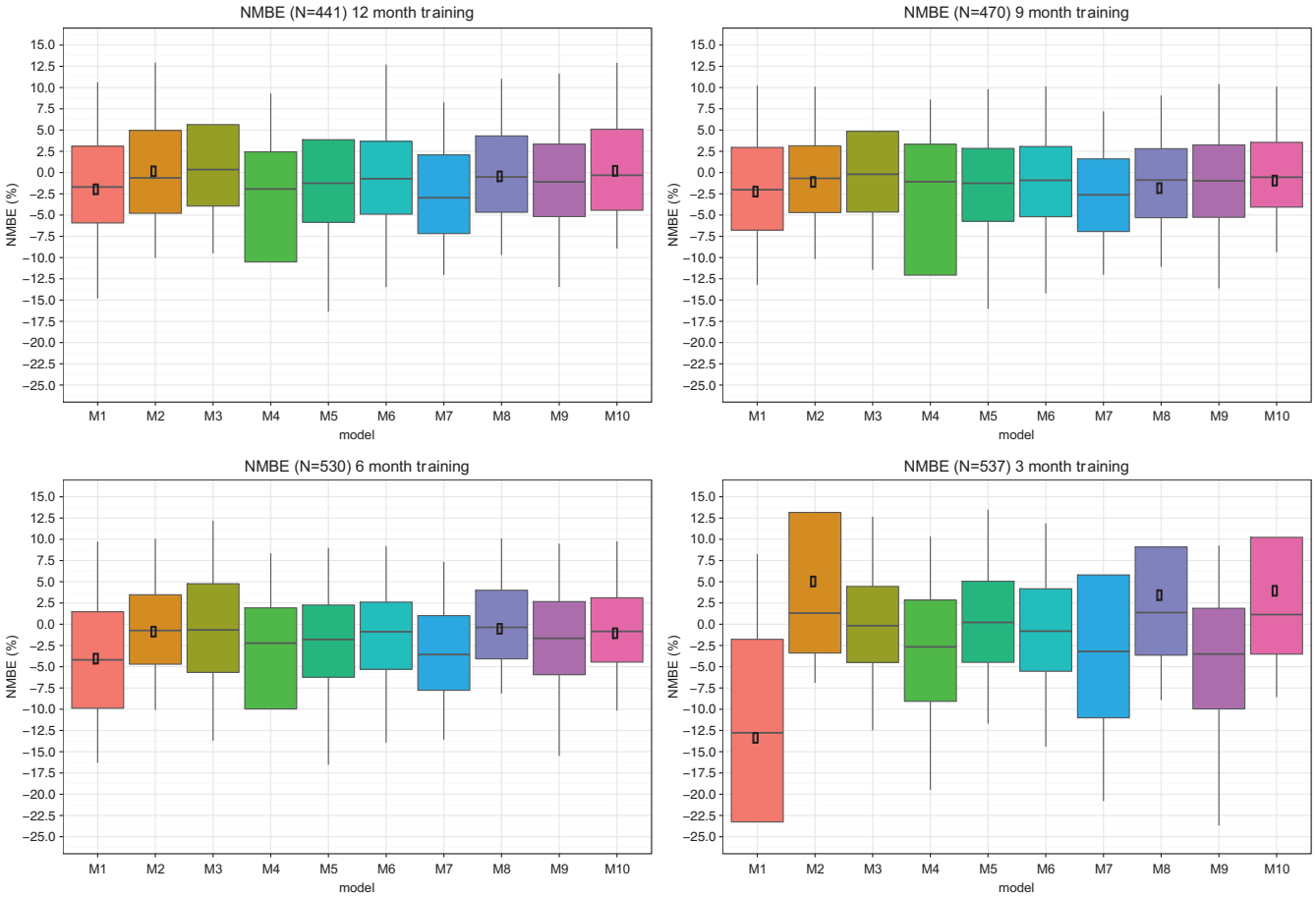

Fig. 4. Distributions of NMBE for each model for a 12-month prediction period, and 12-month, 9-month, 6-month and 3-month training period.

average model NMBE at the 25th, 50th and the 75th percentile (absolute values taken to account for changes in sign), was stable. However, the NMBE increased modestly with six months of training data, and notably with only three months of training data.

3.2. CV(RMSE)

Fig. 5 follows the same conventions as those in Fig. 4, showing distributions of predictive accuracy across the population in the test dataset, for the CV(RMSE) performance metric, calculated for 15-min energy totals. As in Tables 3 and 4 summarizes the 25th, 50th and the 75th percentiles error as the training period is reduced from twelve, to nine, to six, to three months. This provides insight into the general degradation in CV(RMSE) that is seen as the model training period is reduced, while the prediction period is held fixed at twelve months.

Fig. 5 and Table 4 show that when the training period was shortened from twelve months to nine, six, and three months, there was a gradual degradation in predictive accuracy – the average median CV(RMSE) for 15 min energy totals increased from 19.73 to 21.12, 23.54 and 28.58 respectively.

In contrast to the 15-min CV(RMSE) results shown in Figs. 5 and 6 shows the results for the CV(RMSE) performance metric, when calculated for daily energy totals. As expected, errors for the daily CV(RMSE) are smaller than those for the 15-min energy values. Table 5 summarizes just the 25th, 50th and the 75th percentiles error for daily energy totals as the training period is reduced from twelve, to nine, to six, to three months.

Fig. 6 and Table 5 show that when the training period was shortened, there was a gradual degradation in predictive accuracy – the average median CV(RMSE) for daily energy totals increased from 12.93 to 13.76, 15.43 and 20.47 respectively. For the standard whole-building case of twelve months training followed by twelve months of prediction and for all the models except the model 4, which is a very naïve model, the prediction accuracy in term of CV(RMSE) were less than 25 for more than 75% of buildings. For 6 and 9 months of training data, CV(RMSE) for most models was also within 25.

3.3. NMBE vs.CV(RMSE)

Given that stakeholders generally saw value in assessing model performance according to two complementary metrics, it is useful to consider both metrics simultaneously. Fig. 7 shows median NMBE vs. CV(RMSE) for daily energy totals, for a twelve, nine, six and three months training and twelve month prediction period, for each model that was tested. This view into the results allows a comparison of relative model performance, across both metrics. Models that appear closest to the left hand corner between the vertical and the horizontal red lines of the plot are those that minimize both CV(RMSE) and NMBE. For increased clarity the rightmost bound of the *x*-axis corresponding to CV(RMSE) was fixed at 25, which prevented display of Models 1 and 9 from the graph for the 3-month training period (bottom right).

3.4. Results by climate zone

Figs. 8 and 9 shows the results of NMBE and CV(RMSE) for daily energy totals for regions independently, to supplement the

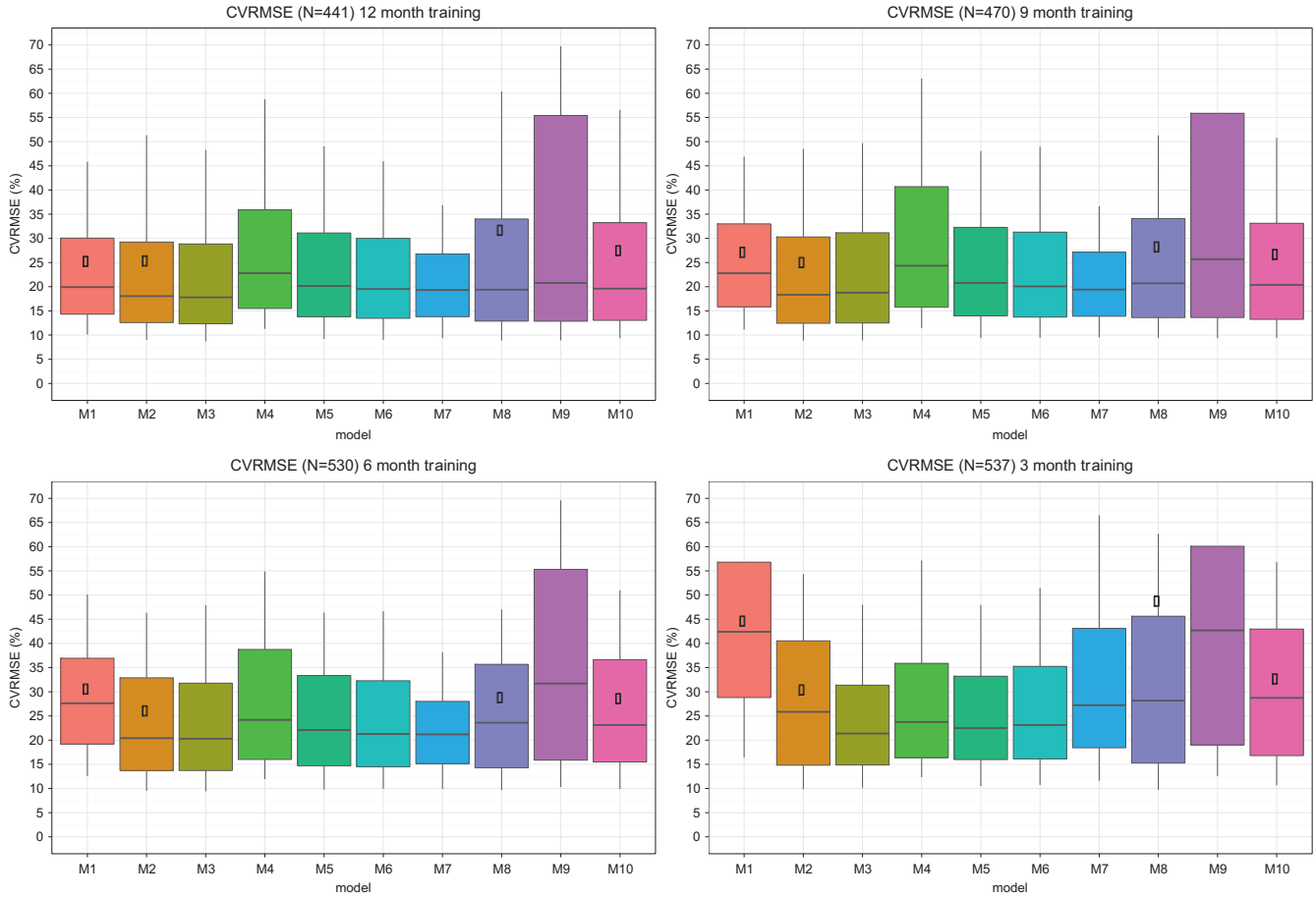

Fig. 5. Distributions of CV(RMSE) for 15-min energy totals for each model, for a 12-month prediction period, and 12-month, 9-month, 6-month, and 3-month training periods.

Table 3
Percentiles of the NMBE for each model, for a 12-month prediction period and 12-month, 9-month, 6-month, and 3-month training periods.

Model	12 months			9 months			6 months			3 months		
	25th	50th	75th	25th	50th	75th	25th	50th	75th	25th	50th	75th
M1	−5.93	−1.7	3.09	−6.78	−2.02	2.95	−9.88	−4.19	1.47	−23.25	−12.77	−1.77
M2	−4.8	−0.63	4.95	−4.71	−0.68	3.15	−4.71	−0.73	3.47	−3.38	1.3	13.16
M3	−3.94	0.35	5.65	−4.65	−0.2	4.85	−5.66	−0.67	4.77	−4.5	−0.17	4.45
M4	−10.51	−1.93	2.43	−12.07	−1.07	3.32	−9.97	−2.22	1.93	−9.07	−2.66	2.85
M5	−5.85	−1.25	3.86	−5.73	−1.26	2.84	−6.23	−1.79	2.26	−4.48	0.21	5.06
M6	−4.9	−0.73	3.67	−5.2	−0.92	3.06	−5.3	−0.88	2.6	−5.54	−0.81	4.17
M7	−7.18	−2.97	2.08	−6.93	−2.62	1.62	−7.77	−3.57	1.02	−11	−3.19	5.81
M8	−4.67	−0.51	4.31	−5.31	−0.88	2.81	−4.07	−0.36	4.01	−3.63	1.38	9.1
M9	−5.18	−1.1	3.35	−5.26	−0.98	3.25	−5.94	−1.65	2.67	−9.96	−3.5	1.88
M10	−4.45	−0.32	5.1	−4.07	−0.55	3.56	−4.46	−0.84	3.12	−3.51	1.14	10.23
Avg. of absolute values	5.74	1.15	3.85	6.07	1.12	3.14	6.4	1.69	2.73	7.83	2.71	5.85

aggregated findings that were detailed in Sections 3.1,3.2 and 3.3. In each plot, distributions of errors across the California dataset are shown in pink and plotted first, those for the Washington, DC dataset are shown in green and plotted second, and those for the Seattle dataset are shown in blue and plotted last. The number of buildings for each analysis time period is shown in the plot title, and the model IDs are displayed in grey across the top of each plot. These plots indicate that regional differences in model performance were observed; the median and the distribution of errors for the California data set ($N = 209$) were modestly smaller than those for the Northwest ($N = 30$), and those for Washington DC ($N = 198$) were notably larger than both California and the Northwest. This may be due to more extreme seasonal variations in outside air temperature in the Mid-Atlantic region. As the California dataset was provided by a participating model developer, while the Northwest and Washington DC datasets were contributed by non-developers, there is also a possibility that the California buildings were less randomly selected from the general commercial stock.

4. Discussion

4.1. Absolute model performance

Overall, the interval data baseline models that were tested were able to predict whole-building energy use with a high degree of

Table 4
Percentiles of the CV(RMSE) for 15-min energy totals for each model, for a 12-month prediction period and 12-month, 9-month, 6-month, and 3-month training periods.

Model	12 months			9 months			6 months			3 months		
	25th	50th	75th	25th	50th	75th	25th	50th	75th	25th	50th	75th
M1	14.32	19.91	30.06	15.84	22.79	33	19.17	27.62	36.94	28.81	42.4	56.8
M2	12.58	18.06	29.22	12.44	18.33	30.29	13.71	20.4	32.88	14.82	25.84	40.52
M3	12.33	17.81	28.84	12.5	18.75	31.17	13.71	20.28	31.78	14.85	21.37	31.37
M4	15.5	22.8	35.89	15.76	24.34	40.68	16	24.18	38.75	16.38	23.74	35.88
M5	13.78	20.16	31.11	13.96	20.78	32.27	14.69	22.09	33.36	15.98	22.49	33.21
M6	13.47	19.53	30.01	13.71	20.06	31.3	14.49	21.26	32.27	16.11	23.12	35.25
M7	13.78	19.32	26.77	13.89	19.41	27.14	15.13	21.19	28.02	18.42	27.22	43.12
M8	12.89	19.39	34	13.61	20.69	34.1	14.28	23.59	35.68	15.26	28.18	45.62
M9	12.87	20.77	55.42	13.62	25.69	55.87	15.85	31.7	55.31	19.01	42.69	60.09
M10	13.04	19.6	33.25	13.24	20.36	33.12	15.47	23.13	36.63	16.8	28.74	42.97
Avg.	13.46	19.73	33.46	13.86	21.12	34.89	15.25	23.54	36.16	17.64	28.58	42.48

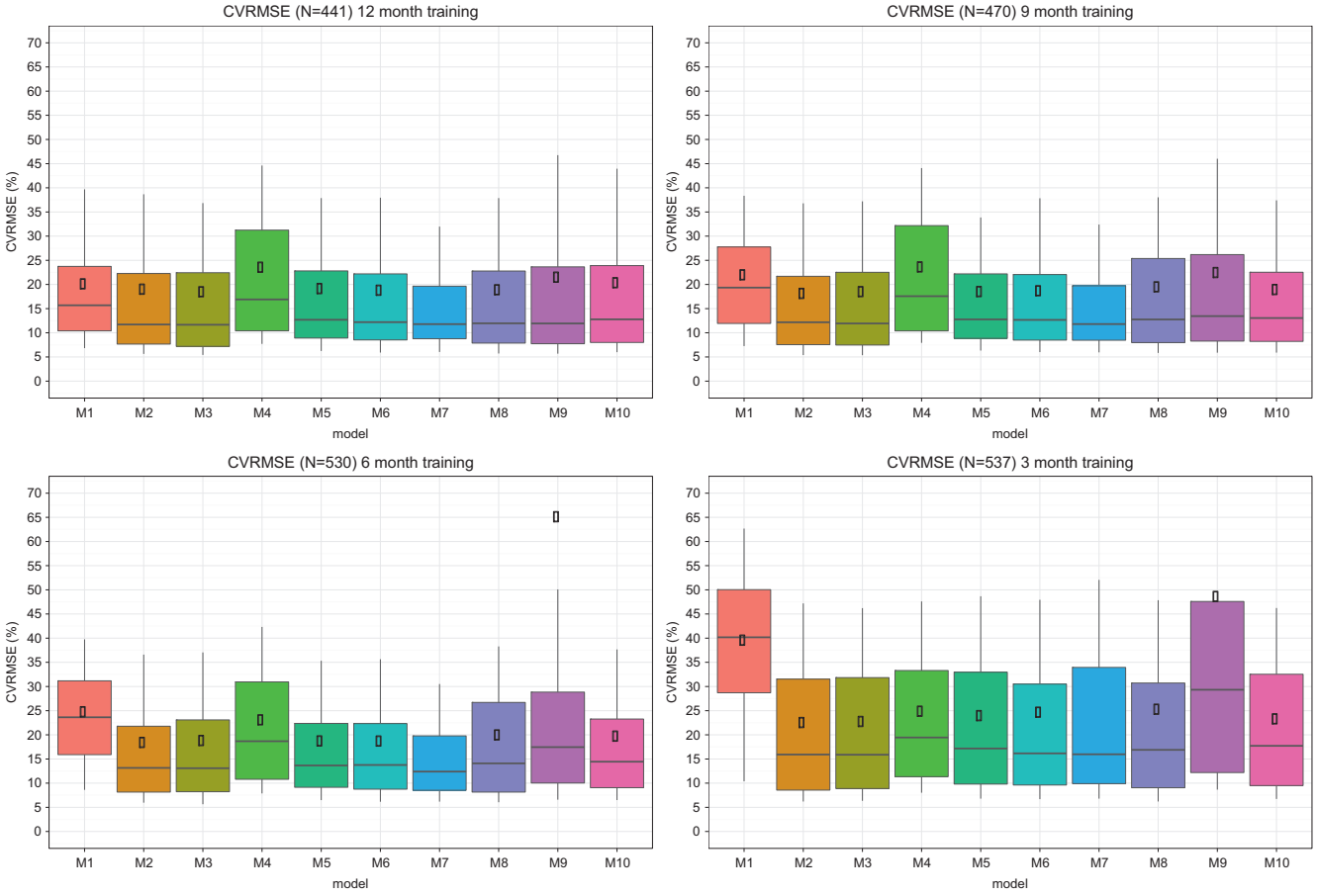

Fig. 6. Distributions of CV(RMSE) for daily energy totals for each model, for a 12-month prediction period, and 12-month, 9-month, 6-month, and 3-month training periods.

accuracy for a large portion of the 537 buildings in the test dataset. For the standard whole-building case of twelve months training followed by twelve months of prediction, and for all models there was a tendency of a bias toward over-predicting energy use (negative NMBE), which has potential implications for pay-for-performance incentive designs. Average CV(RMSE) for daily energy totals was less than 13 for half of the buildings, and less than 24 for three quarters of them (except for model 4, a very naïve, simple model).

This is promising for the industry. ASHRAE Guideline 14 specifies that CV(RMSE) during the *training* period, should be less than 25% if 12 months of post-measure data are used, and no uncertainty analysis is to be conducted [17]. The analyses in this study computed CV(RMSE) during the *prediction* period, which is expected to be even higher than that in the training period. Therefore, while not directly comparable, it appears that the models in this study are likely to meet the ASHRAE requirements for a large fraction of buildings. Median CV(RMSE) for 15-min and daily energy totals was less than 25% for every model tested when twelve months of training data were used. With even six months of training data, median CV(RMSE) for daily energy total was under 25% for all models tested.

Table 5
Percentiles of the CV(RMSE) for daily energy totals for each model, for a 12-month prediction period and 12-month, 9-month, 6-month, and 3-month training periods.

Model	12 months			9 months			6 months			3 months		
	25th	50th	75th	25th	50th	75th	25th	50th	75th	25th	50th	75th
M1	10.4	15.69	23.74	11.94	19.35	27.76	15.89	23.63	31.18	28.7	40.18	50.04
M2	7.66	11.72	22.27	7.54	12.18	21.69	8.15	13.16	21.77	8.56	15.93	31.53
M3	7.19	11.66	22.43	7.47	11.93	22.52	8.22	13.1	23.09	8.88	15.88	31.84
M4	10.41	16.91	31.25	10.39	17.57	32.18	10.8	18.67	30.95	11.34	19.45	33.31
M5	8.92	12.69	22.81	8.85	12.77	22.19	9.14	13.65	22.35	9.81	17.18	32.96
M6	8.52	12.2	22.19	8.48	12.67	22.05	8.75	13.76	22.34	9.62	16.17	30.5
M7	8.78	11.79	19.65	8.46	11.81	19.73	8.53	12.4	19.78	9.9	15.98	33.95
M8	7.87	11.96	22.79	7.98	12.76	25.37	8.15	14.09	26.71	9.05	16.88	30.72
M9	7.73	11.94	23.64	8.27	13.45	26.18	10.03	17.45	28.84	12.19	29.34	47.59
M10	8	12.78	23.91	8.22	13.06	22.53	9.06	14.44	23.27	9.5	17.72	32.53
Avg.	8.55	12.93	23.47	8.76	13.76	24.22	9.67	15.43	25.03	11.76	20.47	35.5

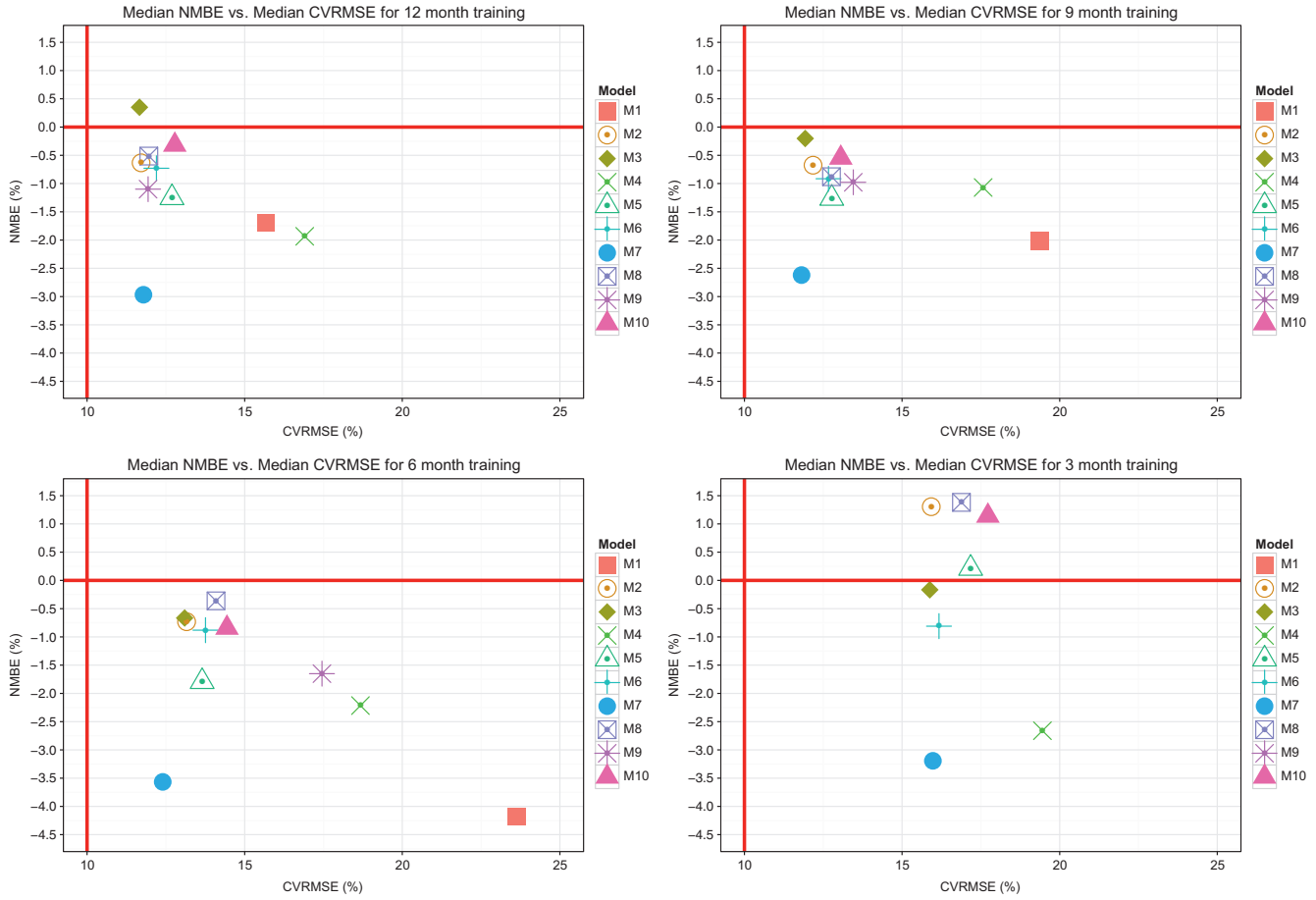

Fig. 7. Median NMBE vs. CV(RMSE) for daily energy totals, for each model tested, a 12-month prediction period, and 12-month, 9-month, 6-month, and 3-month training periods.

Moreover, with NMBE ranging from approximately −1 to 4 for one quarter of the buildings in the dataset, and approximately −1 to −5 for another quarter, the results provide confidence that these M&V approaches will be applicable for many instances of multi-measure programs. This is because *multi-measure* programs commonly target larger savings, on the order of ten percent or more (for example, median retro-commissioning savings are 16% [29]); with errors of just a couple of percent, there is less risk that savings will be 'lost in the noise'. In addition, the accuracies achieved in this study were for a fully automated case. In practice, errors can be further reduced with the oversight of an engineer to conduct non-routine adjustments where necessary. For example, occupancy is not commonly available measured data, and therefore not included in the dataset, or as explanatory variables in the models. Were the buildings to experience significant changes in occupancy, non-routine adjustments might be merited, and could improve the accuracy of the savings that are quantified.

When the training period was shortened from twelve months to nine, and then to six, there was a gradual degradation in predictive accuracy. Not surprisingly, a three-month training period was not long *in general* enough to capture the range of temperatures necessary to reliably predict energy over a the full range of temperatures

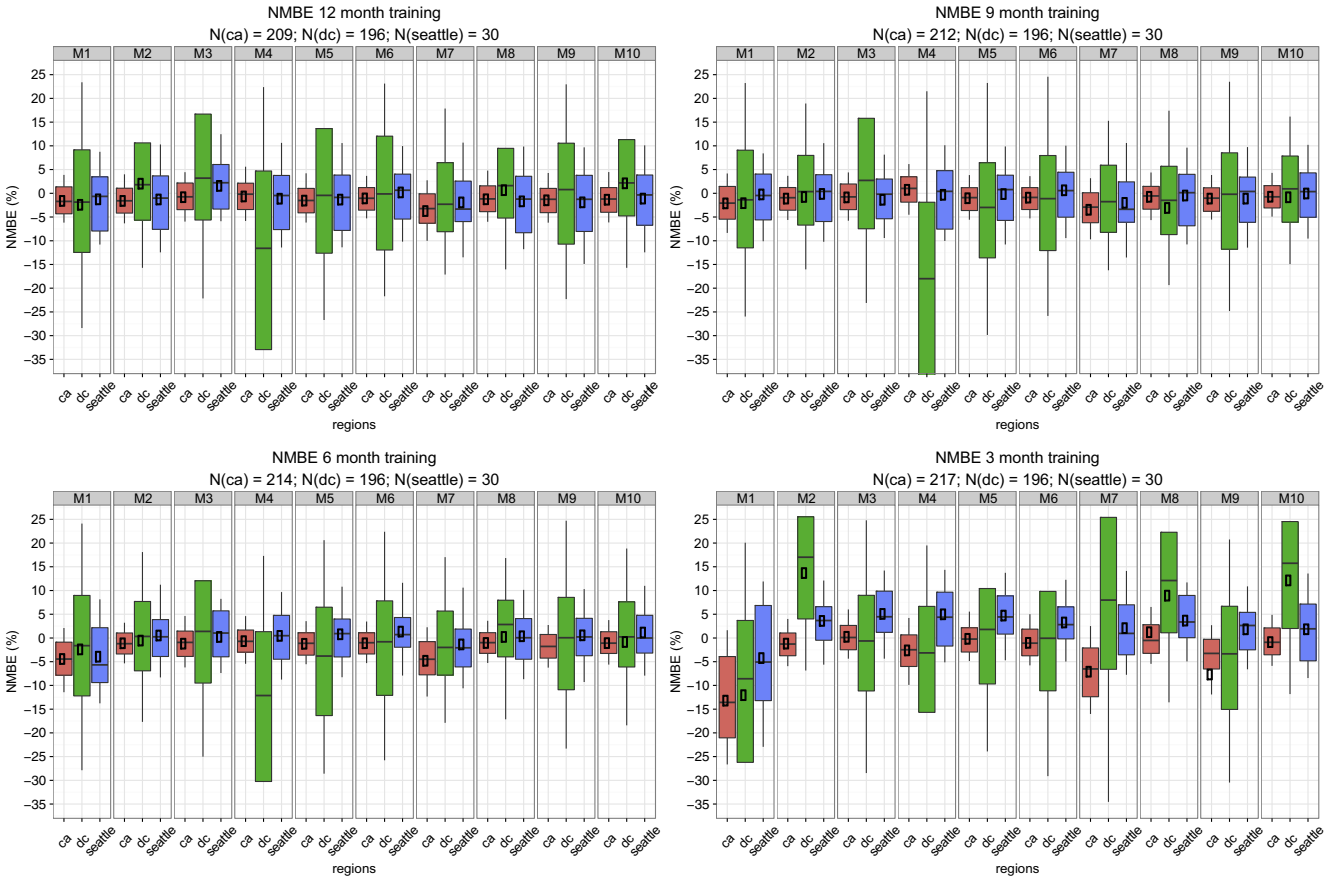

Fig. 8. Distributions of NMBE by climatic region, for each model, for a 12-month prediction period, and 12-month, 9-month, 6-month, and 3-month training periods.

and loads that are seen in a twelve-month period. Given the desire to shorten total time requirements for M&V, the modest increases in error incurred in shortening the training period, in some cases, even to six or three months, may be worth considering in order to reduce the total time necessary to acquire data for the baseline period. This work showed that many of the models that take advantage of increasingly available interval meter data might not require a full 12-months to develop an accurate baseline – even though best practice is commonly accepted to include a full 12 months of pre- and post-data. If the industry begins to more routinely include uncertainty and confidence in reported savings, the effect of reduced time periods (both training and prediction) on reported savings can be analytically quantified.

4.2. Relative model performance

For the most part, each of the ten models performed equally well, according to the two metrics of focus in this study. When plots of median NMBE vs. CV(RMSE) were compared for the standard case of twelve months training and twelve months prediction, Models 1, 4, and 7 emerge as modest outliers; the other models analyzed are relatively tightly clustered together. When non industry-standard shorter training periods (nine, six, and three months) were considered, Models 1, 4, 7, and 9 emerged with relatively higher errors than the other models. However it is important to emphasize that only the median performance was investigated, and in many cases, the magnitude of the difference in errors between models was quite small.

The results section also noted that for some models, the mean error was extremely large. The fact that some buildings are simply not predictable based purely on outside air temperature, date and time is not surprising; there are buildings that are not operated in a predictable manner, for which other drivers of energy use are at play, or for which non-routine adjustments may be appropriate. Interestingly, in some cases the buildings that were poorly predicted by one model, were not the same as the buildings that were predicted poorly by the other models. In addition, most models were unable to generate predictions for at least some of the buildings in the data set – failure rates ranged from roughly zero to ten percent depending on the training period and particular model in question. These aspects of performance are likely due to differences in the underlying form of the models, how they were coded to run automatically in batch mode, their treatment of outliers in the training data, and the different mathematical approaches that they each use.

Table 6 summarizes a qualitative comparative analysis of the tested models in term of prediction accuracy, model complexity and the computational time. The analysis of the prediction accuracy is based on how the models performed in the standard case of twelve months training and twelve months prediction. For the proprietary models, the model complexity summary is based on the model description provided by each model developer. In spite of these relative differences in model performance, it is worth reiterating that absolute performance for all models tested was strong, and provided compelling evidence for their application to whole-building measurement and verification.

5. Conclusions and future work

The results of this work show that interval data baseline models, and streamlining through automation hold great promise for scaling the adoption of whole-building measured savings

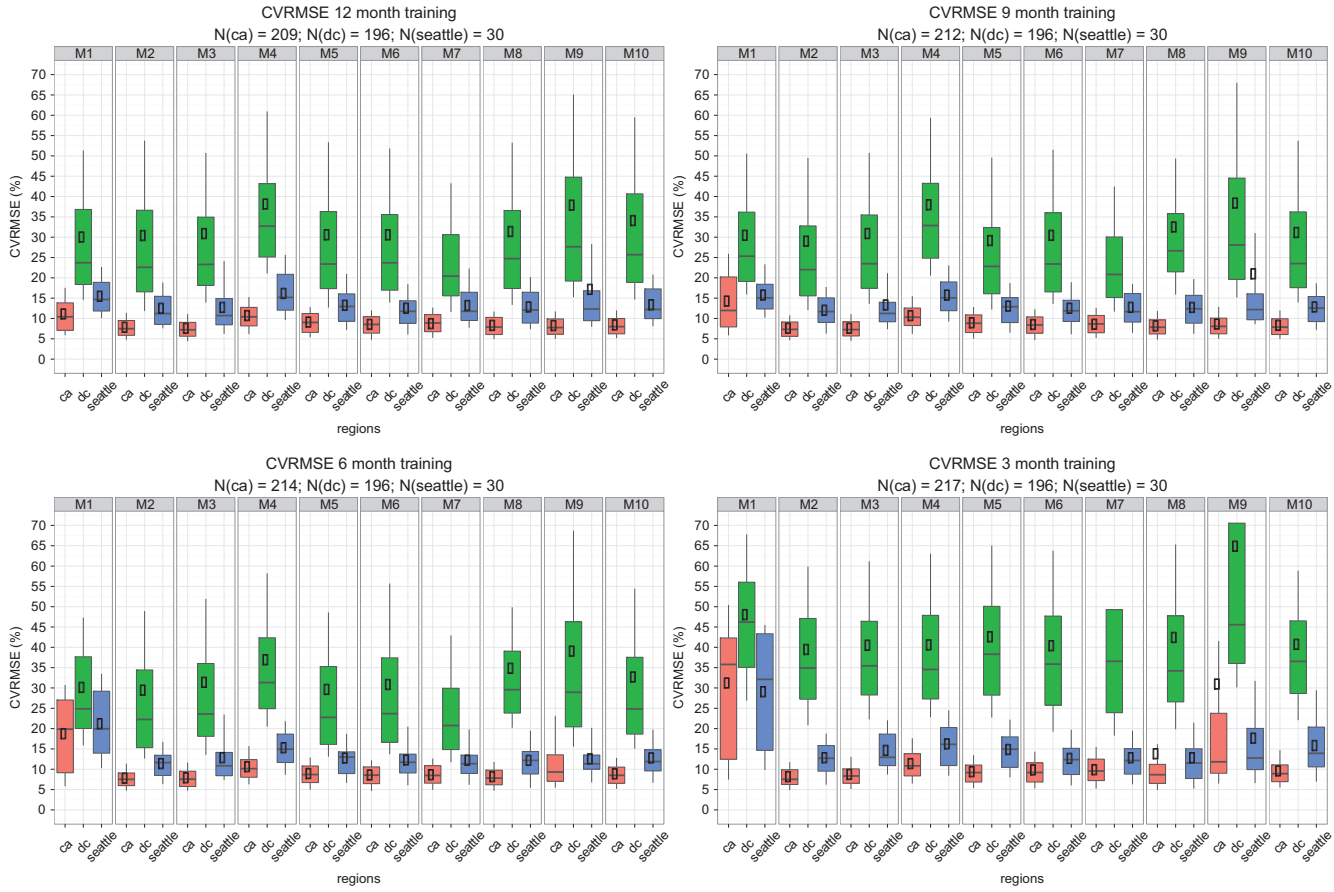

Fig. 9. Distributions of CV(RMSE) for daily energy totals by climatic region, for each model, for a 12-month prediction period, and 12-month, 9-month, 6-month, and 3-month training periods.

Table 6
Comparative summary of the tested models.

Model	Prediction accuracy	Model complexity	Computational time
M1	Medium	Medium–high	Low
M2	High	High	Medium
M3	High	Medium–high	Low
M4	Medium	Low	Low
M5	High	Low	Low
M6	High	Medium	Medium
M7	Medium	Medium–high	Medium–high
M8	High	High	High
M9	High	Medium–high	Low
M10	High	Medium–high	High

calculations using Advanced Metering Infrastructure (AMI) data. These findings can be used in a number of ways, by diverse stakeholders in M&V and in the delivery of efficiency programs. First, they can be used by program administrators to qualify technologies for readiness, and to pilot whole-building, and pay-for-performance programs that require metered whole building energy savings estimation. The test procedures and metrics can be replicated for additional models of interest, and in cases where territory-specific results are desired a local test data set can be used. The results can also be used to pre-vet M&V plans for specific projects, given project requirements for uncertainty in reported savings. Conducted in collaboration with regulators and evaluators, this can increase the transparency in the M&V process, and the reliability of savings estimation.

While uncertainty is not commonly considered today, it could hold value for evaluating and reducing project and investment risk. For example, ASHRAE's published methods for computing fractional savings uncertainty depend on depth of savings, length of the training and prediction periods, and model CV(RMSE). "Look-up" tables based on results such as those presented in this study can be used to explore the likelihood that a given model will produce savings estimations that meet uncertainty and confidence requirements, for a specific set of buildings and expected depth of savings. After an efficiency project is initiated, these methods can be used as the project progresses to track achieved savings relative to expected savings, and perhaps even be used to indicate when measures are not correctly implemented, or when non-routine changes have occurred in the building operations or loads.

Future work will focus on four key areas: (1) application of these automated approaches in partnership with utilities and implementers, using data from buildings that have participated in programs or pilots; (2) exploration of industry demand for the objective model testing methods as presented in this paper, and identification of appropriate bodies to which the procedures should be transferred; (3) continued engagement of the evaluator, program manager and implementer community to collectively more clearly define uncertainty and confidence requirements for reporting gross energy savings; (4) research to explore the most appropriate methods of calculating the uncertainty associated with savings calculations.

Acknowledgements

This work was supported by the Assistant Secretary for Energy Efficiency and Renewable Energy, Building Technologies Program,

of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

The authors would like to thank the members of the project Technical Advisory Group for their participation and feedback throughout the course of the work. The authors also acknowledge each of the developers who submitted baseline models for inclusion in this study. Those who chose to self-identify include Buildings Alive Pty. Ltd. of Sydney Australia, Paul Raftery and Tyler Hoyt of UC Berkeley's Center for the Built Environment, Gridium Inc., Lucid Design Group, and Performance Systems Development of New York, LLC. Throughout the project, Cody Taylor of the Building Technologies Office provided valuable support and guidance. Finally we would like to thank those who contributed to the test dataset. Without a sufficient volume and diversity of data, meaningful insights would not have been possible.

Appendix A

A.1. Description of models

For cases in which the model developer consented, more detailed descriptions of the baseline models are provided below.

M4: Mean week

In this model the predictions of the future values, for a given day of the week d and time t, are equal to the average of the training data for this particular day and time, then we can write the predictions as

$$\hat{y}(d,t) = \frac{1}{N(d,t)} \sum_{i=1}^{N(d,t)} y(d_i, t_i)$$

where $y(d_i, t_i)$ is the value of the ith week of the training data, and $N(d,t)$ is the number of weeks in the training data, which have values for the day of the week d and time t.

M5: Time of week and temperature

In the *Time of Week and Temperature* model, the predicted load is a sum of two terms: (1) a "time of week effect" that allows each time of the week to have a different predicted load from the others, and (2) a piecewise-continuous effect of temperature. The temperature effect is estimated separately for periods of the day with high and low load, to capture different temperature slopes for occupied and unoccupied building modes. The model is described in detail in [11].

For each day of the week, the 10th and 90th percentile of the load were calculated; call these L_{10} and L_{90}. The first time of that day at which the load usually exceeds the $L_{10} + 0.1 * (L_{90} - L_{10})$ is defined as the start of the "occupied" period for that day of the week, and the first time at which it usually falls below that level later in the day is defined as the end of the "occupied" period for that day of the week.

M6: Weighted time of week and temperature

This model is the *Time-of-Week and-Temperature* model with the addition of a weighting factor to give more statistical weight to days that are nearby to the day being predicted. This is achieved by fitting the regression model using weights that fall off as a function of time in both directions from a central day. In the implementation used in this work, the weight parameter is set to fourteen, placing more weight on the most recent two weeks of data.

M7. Ensemble approach combining nearest neighbors and a generalized linear model, developed by Lucid Design Group.

Lucid's model employed a sequential ensemble approach, first generating predictions using K-nearest-neighbors (KNN), and then adjusting the KNN output with help of a ridge regression model. The intuition underlying this approach is that KNN is generally strong in capturing nonlinearities in the relationship between prediction and outcome variables, especially for low-dimensional problems. However, its applicability is bounded by the availability of sufficiently "nearby" neighbors for each prediction made.

In an M&V context, this problem might manifest as negative bias when predicting demand on hot days, especially if the training set spans a period of mostly cooler temperatures, because of either seasonal or year-to-year variation. This limitation is addressed by adjusting each KNN prediction to account for this potential bias. The prediction process works as follows:

First, a linear model of the following form is built, minimizing least squares, and using ridge regularization penalty, tuned using leave-one-out (LOO) cross-validation.

$$y_{d,i} = \alpha_0 + \alpha_1\ Temp_{d,i} + \alpha_1\ TA_{d,i} + \alpha_2\ TB_{d,i} + \sum_{n=1}^{96} \alpha_{2+n} I[i = n]$$

where *Temp* refers to temperature in Fahrenheit, *TA* and *TB* are transformed temperature variables as defined below, and the remaining terms are indicator variables for each of 96 quarter hour periods in a day.

$$TA_{d,i} = \begin{cases} 0 & T_t < 65^\circ\mathrm{F} \\ (T_{d,i} - 65^\circ\mathrm{F})^2 & T_t \geqslant 65^\circ\mathrm{F} \end{cases}$$

$$TB_{d,i} = \begin{cases} 0 & T_t \geqslant 65^\circ\mathrm{F} \\ (T_{d,i} - 65^\circ\mathrm{F})^2 & T_t < 65^o F \end{cases}$$

Then five representative features of each 24-h period in the training set are calculated. The notion of distance between days is defined to be a weighted Euclidian distance in the resulting $\mathbb{R}^5$ coordinate space.

Features	Weights
Maximum daily temperature	1.0
Minimum daily temperature	1.0
Business day (binary indicator)	2.0
Winter seasonal factor	0.5
Summer seasonal factor	0.5

To derive the summer and winter seasonal factors, picked two "diametrically opposed" dates are picked – January 15th and July 15th – to represent the midpoint of the summer and winter seasons from a meteorological standpoint. Then a value in the range [0,1] is calculated to represent the distance from that seasonal midpoint for each date. Here, *DayDelta* refers to the absolute difference between two dates, measured in days.

$$Summer\ Factor_d = \frac{\max(0, 90 - DayDelta(Date_d,\ July\ 15th))}{90}$$

$$Winter\ Factor_d = \frac{\max(0, 90 - DayDelta(Date_d,\ January\ 15th))}{90}$$

Predictions are then made one day at a time, in two phases. In the first phase, a KNN is used to approach to select K similar days, where K is the lesser of 15% and 20% of the number of days available in the training set. The demand from those K days is combined interval-by-interval using a weighted average, where the weight for each day decreases with increasing distance from the day being predicted.

$$weight_d \propto \frac{1}{1 + Distance(Date_{prediction}, Date_d)}$$

The output of this step is 96 values $\hat{y}_{d,i}$ predicting demand each quarter hour interval i of the day d being predicted. The second and final step is to adjust that result using our linear model from the first step. To do that, we use our linear model to predict demand $\hat{r}_{d,i}$ for each interval i of each day d and in our set of nearest neighbors. Note that the same linear model is used to predict demand for the day being predicted.

Finally, the interval-by-interval difference between the nearest neighbor predictions and the target day prediction is taken, and the KNN output is adjusted by those differences to generate a final prediction:

$$\hat{y}_{d,i} = \hat{q}_{d,i} - 0.6 \times \sum_{d=1}^{K} weight_d \left(\hat{r}_{d,i} - \hat{r}_{prediction,i}\right)$$

The 0.6 factor is inserted because it was found that applying the full adjustment overcompensated for the local biases of KNN alone, and reduced the RMSE in cross validation trials. Future improvements on this approach might attempt to tune that value as a parameter rather than use a "magic number."

References

[1] Energy Information Administration. Commercial buildings energy consumption survey (CBECS). U.S. Department of Energy; 2003.
[2] Consortium for Energy Efficiency (CEE). 2013 state of the efficiency program industry: budgets, expenditure, and impacts; 2014.
[3] Satchwell A, Goldman C, Larsen P, Gilligan D, Singer T. A survey of the US ESCO industry: market growth and development from 2008 to 2011. Lawrence Berkeley National Laboratory. Report Number LBNL-3479E; 2010.
[4] Jayaweera T, Haeri H, Kurnik C. The uniform methods project: methods for determining energy efficiency savings for specific measures. national renewable energy laboratory. April 2013. NREL Report # NREL/SR-7A30-53827; 2013.
[5] Granderson J, Piette MA, Ghatikar G. Building energy information systems: user case studies. Energ Efficiency 2011;4(1):17–30.
[6] Kramer H, Russell J, Crowe E, Effinger J. Inventory of commercial energy management and information systems (EMIS) for M&V applications, prepared by PECI for northwest energy efficiency alliance. Report Number E13-264; 2013.
[7] Rogers Ethan A, Carley Edward, Deo Sagar, Grossberg Frederick. How information and communications technologies will change the evaluation, measurement, and verification of energy efficiency programs; 2015.
[8] Northeast Energy Efficiency Partnerships. The changing EM&V paradigm: a review of key trends and new industry developments, and their implications on current and future EM&V practices. American council for an energy efficient economy research report IE1503; 2015.
[9] Energy and Environmental Economics. Time dependent valuation of energy for developing building efficiency standards. Report prepared for the California Energy Commission; 2011.
[10] Granderson J, Addy N, Price P, Sohn M. Automated measurement and verification: performance of public domain whole-building electric baseline models. Appl Energy 2015;144:106–13.
[11] Mathieu JL, Price PN, Kiliccote S, Piette MA. Quantifying changes in building electricity use, with application to Demand Response. IEEE Trans Smart Grid 2011;2:507–18.
[12] Piette MA, Brown RE, Price PN, Page J, Granderson J, Riess D, et al. Automated measurement and signaling systems for the transactional network. Lawrence Berkeley National Laboratory; 2013 [December 2013. LBNL-6611E].
[13] Srivastav A, Tewari A, Dong B. Baseline building energy modeling and localized uncertainty quantification using Gaussian mixture models. Energy Build 2013;65:438–47.
[14] Heo Y, Zavala VM. Gaussian process modeling for measurement and verification of building energy savings. Energy Build 2012;53:7–18.
[15] Zhao HX, Magoulès F. A review on the prediction of building energy consumption. Renew Sustain Energy Rev 2012;16(6):3586–92.
[16] Efficiency Valuation Organization (EVO). International performance measurement and verification protocol: concepts and options for determining energy and water savings. 1, EVO 10000-1:2012; 2012.
[17] ASHRAE Guideline 14. ASHRAE guideline 14-2014 for measurement of energy and demand savings, american society of heating, refrigeration and air conditioning engineers. Atlanta, GA; 2014.
[18] Granderson J, Price PN. Development and application of a statistical methodology to evaluate the predictive accuracy of building energy baseline models. Energy 2014;66:981–90.
[19] Price P, Jump D, Sohn M. Functional testing protocols for commercial building efficiency baseline modeling software. Prepared by LBNL and QuEST for pacific gas and electric. Report number LBNL-6593E, Pacific Gas and Electric Report Number ET12PGE5312; 2013.
[20] Walter T, Price PN, Sohn MD. Uncertainty estimation improved energy measurement and verification procedures. Appl Energy 2014;130:230–6.
[21] Haberl JS, Thamilseran S. The great energy predictor shootout II: measuring retrofit savings. ASHRAE J 1998;40(1):49–56.
[22] Kreider JF, Haberl JS. Predicting hourly building energy usage. ASHRAE J 1994;36(6):1104–18.
[23] Baechler M, Williamson J, Gilbride T, Cole P, Hefty M, Love PM. Building America best practice series, Volume 7.1. High performance home technologies: guide to determining climate regions by county. Prepared for by Pacific Northwest National Laboratory, and Oakridge National Laboratory, August 2010, Report Number PNNL-17211; 2010.
[24] Breiman L. Random forests. Mach Learn 2001;45(1):5–32.
[25] Geurts P, Ernst D, Wehenkel L. Extremely randomized trees. Mach Learn 2006;63(1):3–42.
[26] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2009.
[27] McCullagh P, Nelder JA. Generalized linear models, vol. 37. CRC Press; 1989.
[28] Friedman JH. Multivariate adaptive regression splines. Ann Statis 1991:1–67.
[29] Mills E. Building commissioning: a golden opportunity for reducing energy costs and greenhouse gas emissions in the United States. Energy Efficiency 2011;4(2):145–73.