M&V Shootout: Setting the Stage for Testing the Performance of New Energy Baseline Models

Dr. Jessica Granderson, Lawrence Berkeley National Laboratory, Berkeley CA
Dr. Samir Touzani, Lawrence Berkeley National Laboratory, Berkeley CA
Claudine Custodio, Lawrence Berkeley National Laboratory, Berkeley CA
Dr. Michael Sohn, Lawrence Berkeley National Laboratory, Berkeley CA
Samuel Fernandes, Lawrence Berkeley National Laboratory, Berkeley CA
Dr. David Jump, Quantum Energy Services and Technologies, Berkeley CA
Cody Taylor, Department of Energy Building Technologies Office, Washington, DC

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 data acquisition and often do not deliver results until years after the program period has ended. A spectrum of savings calculation approaches are used, with some relying more heavily on measured data and others relying more heavily on estimated or modeled data, or stipulated information.

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. Energy management and information systems (EMIS) technologies, not only enable significant site energy savings, but are also beginning to offer M&V capabilities. This paper expands recent analyses of public-domain, whole-building M&V methods, focusing on more novel baseline modeling approaches that leverage interval meter data. We detail a testing procedure and metrics to assess the performance of these new approaches using a large test dataset. We also provide conclusions regarding the accuracy, cost, and time trade-offs between more traditional M&V and these emerging streamlined methods. Finally, we discuss the potential evolution of M&V to better support the energy efficiency industry through low-cost approaches, and the long-term agenda for validation of building energy analytics.

Introduction

Measurement and verification (M&V) of energy efficiency measures can be critical to establishing the value of efficiency both to building owners and to utility programs incentivizing savings. However, M&V can be costly and time consuming. Depending on the M&V methods used and whether third party evaluation is included, M&V costs can range from 1-5% of project portfolio costs (Jayaweera et al. 2013). Today, the growing availability of data from smart meters and devices, combined with time series data analytics offers the potential to streamline the M&V process through increased levels of automation. In addition, energy management and information systems (EMIS)[1] are beginning to automatically create

[1] Energy Management and Information Systems (EMIS) are software tools that store, analyze, and display energy use or building systems data; they include energy information systems, and fault detection and diagnostics systems (Kramer et al. 2013).

baseline models and calculate energy savings according to the principles of the International Performance Measurement and Verification Protocol (Kramer et al. 2013). These technologies offer the potential to reduce the time and costs necessary to conduct M&V.

Although these emerging technologies and approaches hold great promise to scale M&V in the commercial buildings sector, several questions relating to their use remain to be answered: How accurate are automated baseline models that utilize interval meter data? How can proprietary tools that automate gross savings calculations be evaluated? How can one tool or model be compared to another? What metrics should be used to quantify the performance of these tools?

In this paper, we begin to answer these questions by presenting 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 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. The analyses that are presented 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.

Methodology

The evaluation of model predictive accuracy that is presented in this paper is based on a 4-step testing procedure, generally characterized as statistical cross validation using large test datasets. This procedure is depicted in Figure 1. The test dataset comprises interval meter data and independent variable data, such as outside air temperature, for dozens to 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. Figure 2 shows an example of actual, and model-predicted data for a 12-month training period and a 12-month prediction period. Performance metrics that quantify the difference between the model prediction and the actual load are calculated and used to characterize accuracy.

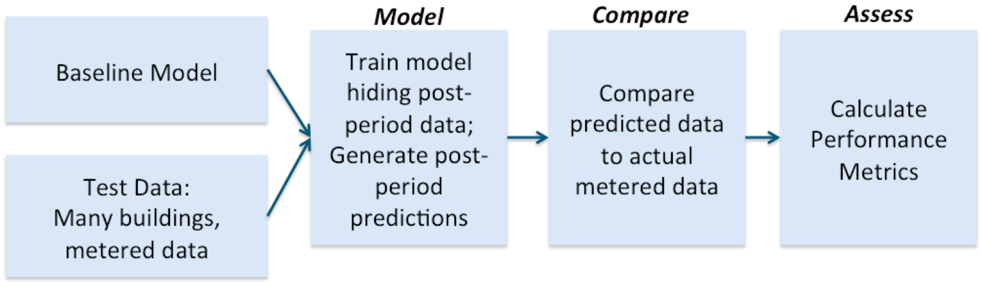
Model
Compare
Assess
Baseline Model
Test Data: Many buildings, metered data
Train model hiding post-period data; Generate post-period predictions
Compare predicted data to actual metered data
Calculate Performance Metrics

Figure 1. Schematic of the test procedure used to evaluate the performance of automated M&V methods.

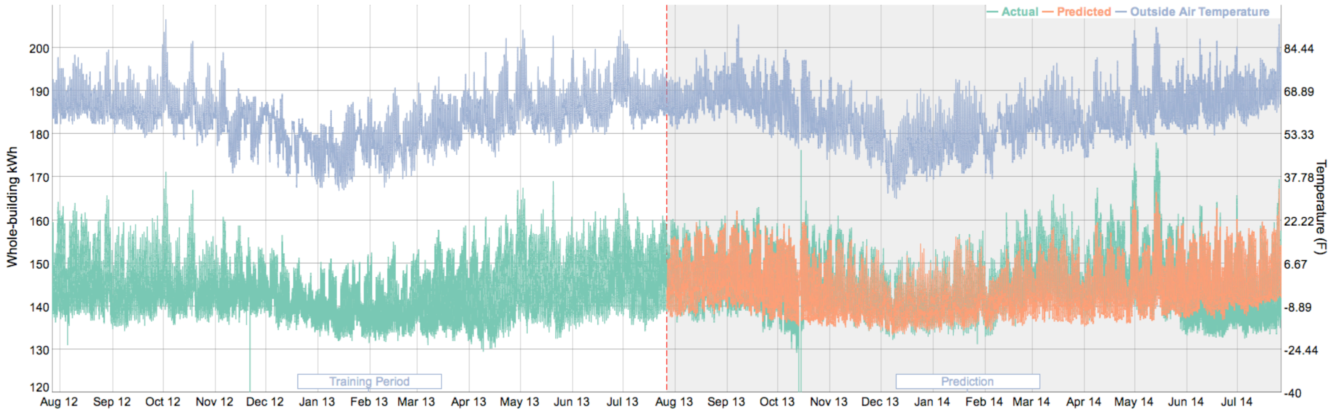

Figure 2. Actual and model-predicted energy data, overlaid with outside air temperature, for a 12-month training period and 12-month prediction period.

This test procedure is documented in further detail in prior publications (Granderson et al. 2015; Granderson et al. 2014). It shares important similarities to the approaches used in the ASHRAE ‘shootouts’ of the mid and late 1990s (Haberl & Thamilseran 1998; Kreider & Haberl 1994). In both cases, cross-validation is used to determine model error, and in both cases, similar performance metrics are considered. However, the ASHRAE shootouts were limited to data from a total of three buildings, and the cross-validation was conducted during a short subset of the model *training* period. The ASHRAE competitions considered total energy use from a sum of submetered quantities, but the analyses presented in this work are constrained to data and models of whole building electric metering because that is the only meter data that was available in our dataset; it is also the type of interval data most readily available in today’s buildings.

Test Data

The test dataset for the analyses presented in this paper comprised 537 commercial buildings from multiple climate zones, and is characterized in Table 1. These data represent a combined pool from model developers, a utility, and a municipality. For each building, 15-minute whole-building electricity data was paired with zip-code based data for outside air temperature. Buildings in Climate Zone 3 were from Northern and Central California, and those from Climate Zone 4 were from the Northwest and Mid-Atlantic.

Table 1. Distribution of buildings in the test dataset based on ASHRAE climate zones

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

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 are described below, with references and a description for those that are published in the literature. While the models may be able to accommodate additional independent variables were they available, 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. The ten models are:

M1. *Combination principle component analysis and bin modeling,* developed by Buildings Alive Pty. Ltd. of Sydney Australia.

M2. *Combination Random Forest, Extra-Trees (extremely randomized trees) 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* (Granderson et al. 2015): 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* (Piette et al. 2013): 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 and a generalized linear model*, developed by Lucid Design Group.

M8. *Combination Multivariate Adaptive Regression Splines (MARS) and other advanced regression*

M9. *Combination bin modeling and other advanced regression,* developed by Performance Systems Development of Ney York, LLC.

M10. *Nearest neighbor advanced regression*

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, the most meaningful two metrics were found to be normalized mean bias error (NMBE), and coefficient of variation of the root mean squared error (CV(RMSE)). These two metrics provide a nice complement in understanding model performance for M&V applications. NMBE gives a sense of the total difference between model predicted energy use and actual metered energy use, with intuitive implications for the accuracy of avoided energy use calculations. CV(RMSE), 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 prominent in resources such as ASHRAE Guideline 14. These metrics are defined in 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. In the case of CV(RMSE), results are presented for daily totals of energy use across the prediction period; for NMBE, by definition, the metric captures the percent error in measured versus predicted energy use for the full prediction period.

$$Normalized\,Mean\,Bias\,Error = \frac{\frac{1}{N}\sum_{i}^{N}\left(y_i - \hat{y}_i\right)}{\bar{y}} \times 100$$

$$CV\,Root\,Mean\,Squared\,Error = \frac{\sqrt{\frac{1}{N}\sum_{i}^{N}\left(y_i - \hat{y}_i\right)^2}}{\bar{y}} \times 100$$

Time Horizons

In keeping with current standard practice and guidelines for whole-building avoided energy use calculations (ASHRAE 2012), the analyses in this study are grounded in a 12-month 'post', or model prediction period. We assess the degradation in prediction accuracy when the 'pre', or model training period is reduced from 12-months, to shorter and shorter time horizons. Specifically, results are presented for 12-month, 9-month, 6-month, and 3-month training periods.

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.

Normalized Mean Bias Error

Normalized mean bias error across the full population of buildings in the test dataset is shown for each model, in Figure 2. 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, or approaches infinity, and therefore is 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 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 is shown in the title at the top of each plot. Most models that were tested failed to generate predictions for at least some of the buildings in the test dataset.

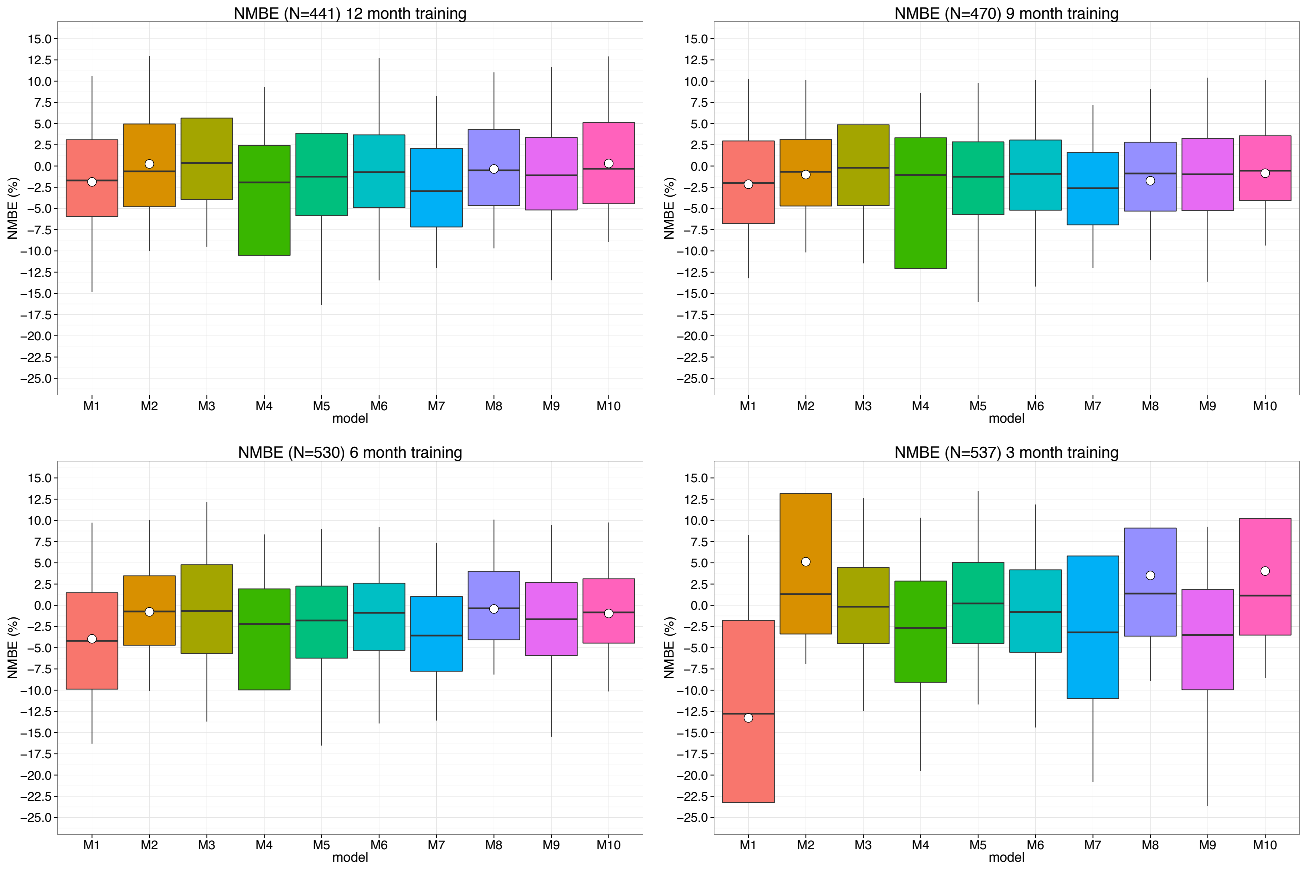

Figure 3. Distributions of NMBE for each model, for a 12-month prediction period, and 12-month, 9-month, 6-month, and 3-month training periods.

While Figure 3 shows percentiles of errors across the full population of buildings and training periods that were analyzed, Table 2 summarizes just the median, 50th percentile 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 Figure 3 and Table 2 show that for the majority of models there was a tendency of a bias toward over-predicting the energy use (NMBE negative), In addition, when the training period was shortened from twelve months to nine, the average model NMBE (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.

Table 2. Median NMBE for each model, for a 12-month prediction period and 12-month, 9-month, 6-month, and 3-month training periods.

Model	Model Training Period			
	12 months	9 months	6 months	3 months
M1	-1.7	-2.02	-4.19	-12.77
M2	-0.63	-0.68	-0.73	1.3
M3	0.35	-0.2	-0.67	-0.17
M4	-1.93	-1.07	-2.22	-2.66
M5	-1.25	-1.26	-1.79	0.21
M6	-0.73	-0.92	-0.88	-0.81
M7	-2.97	-2.62	-3.57	-3.19
M8	-0.51	-0.88	-0.36	1.38
M9	-1.1	-0.98	-1.65	-3.5
M10	-0.32	-0.55	-0.84	1.14
Avg. of Absolute Median Values	1.15	1.12	1.69	2.71

CV(RMSE), Daily Energy Totals

Figure 4 shows the results for the CV(RMSE) performance metric, when calculated for daily energy totals across each day in the prediction period. Similar to Table 2, Table 3 summarizes just the median, or 50th percentile error as the training period is reduced from twelve, to nine, to six, to three months. The results in Figure 4 and Table 3 show that when the training period was shortened, there was a gradual degradation in predictive accuracy. For these training periods the average median CV(RMSE) for 15min 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 CV(RMSE) was less than 25 for more than 75% of buildings.

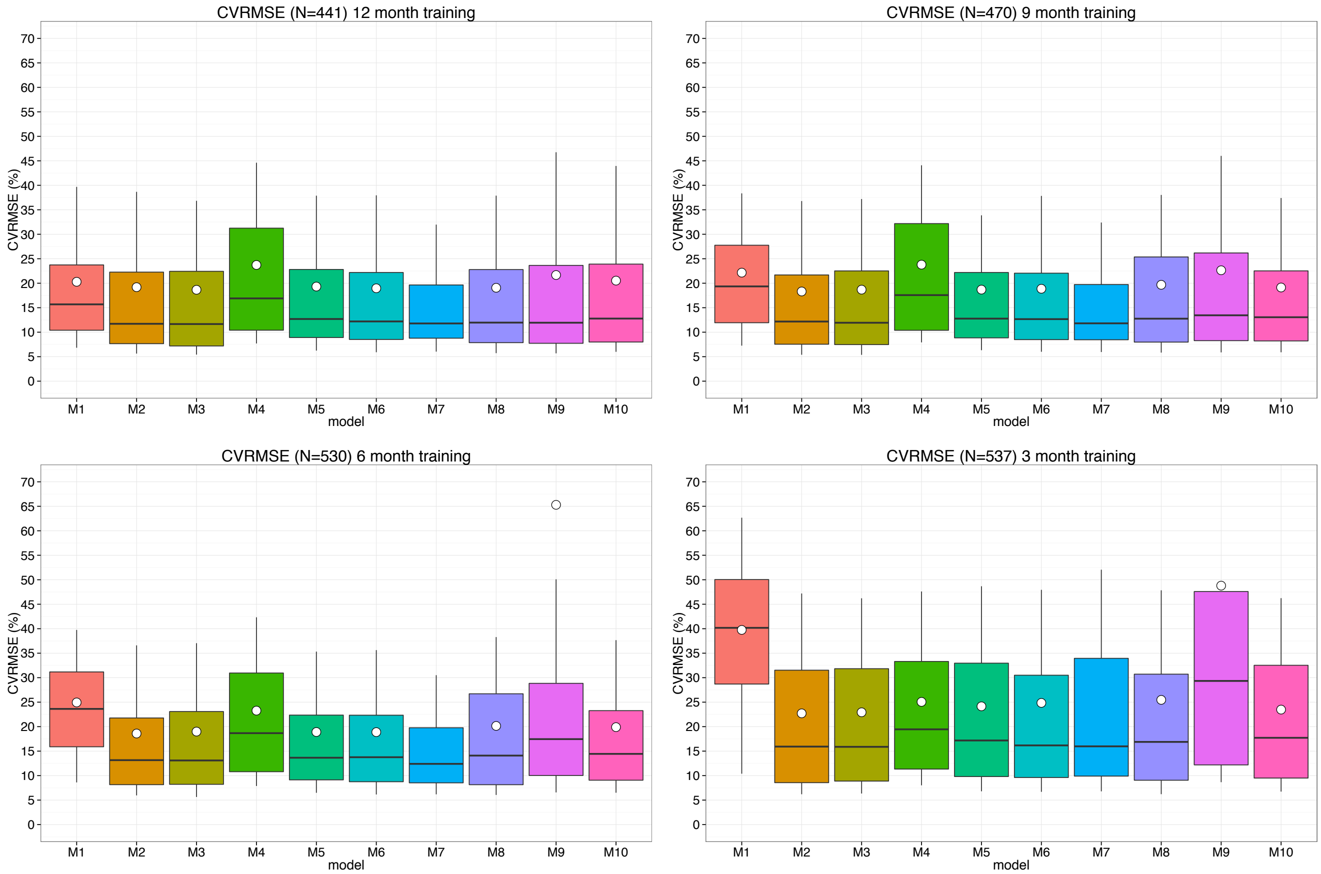

Figure 4. Distributions of CV(RMSE) for daily energy totals for each model, for a 12-month prediction period, and 12-month, 9-month, 6-month, and 3-month training periods.

Table 3. Median 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	**Model Training Period**			
	12 months	**9 months**	**6 months**	**3 months**
M1	15.69	19.35	23.63	40.18
M2	11.72	12.18	13.16	15.93
M3	11.66	11.93	13.1	15.88
M4	16.91	17.57	18.67	19.45
M5	12.69	12.77	13.65	17.18
M6	12.2	12.67	13.76	16.17
M7	11.79	11.81	12.4	15.98
M8	11.96	12.76	14.09	16.88
M9	11.94	13.45	17.45	29.34
M10	12.78	13.06	14.44	17.72
Average	12.93	13.76	15.43	20.47

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. Figure 5 shows median NMBE vs. CV(RMSE) for daily energy totals, for a 12-month training and prediction period, for each model that was tested. Although not shown here, these plots were generated and analyzed for the nine-month, six-month, and 3-month training periods as well. This view into the results allows a comparison of relative model performance, across both metrics. Models that appear closest to the upper left hand corner of the plot between the vertical and the horizontal red lines are those that minimize both CV(RMSE) and NMBE.

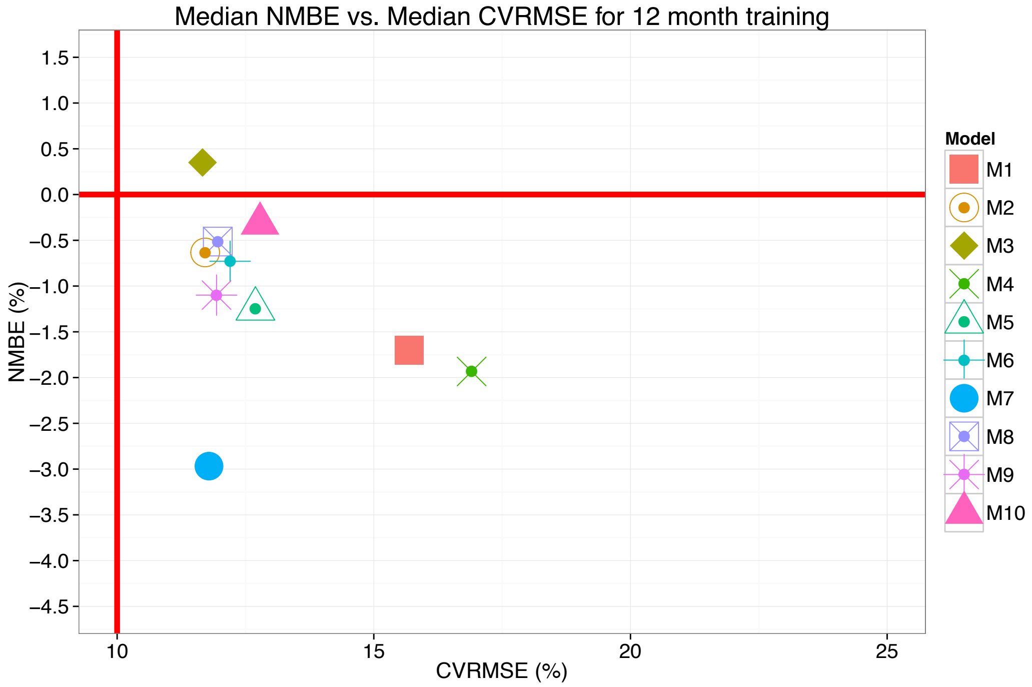

Figure 5. Median NMBE vs. CV(RMSE) for daily energy totals, for each model tested, for a twelve-month training period and twelve-month prediction period.

Discussion

Absolute Model Performance

Overall, the interval data models that were tested were able to predict whole-building energy use with a high degree of 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 the energy use (negative NMBE), which has potential implications for pay-for-performance incentive designs. Average CV(RMSE) for daily energy totals 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 (ASHRAE 2002). 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 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.

Moreover, with NMBE ranging from approximately -1 to 4 for one quarter of the buildings in the data set, and approximately -1 to -5 for another quarter of the buildings, 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% (Mills 2011)); 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 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 reducing the total time necessary to acquire data for the baseline period.

Climatic Differences

The test dataset that was compiled for this analysis comprised whole-building data that represented a dataset of convenience, as opposed to design. Ideally, the buildings would be uniformly distributed across all climate zones, 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 (ASHRAE Climate Zone 3), and Washington, DC (ASHRAE Climate Zone 4), with much less representation from other climates. Although not presented in detail in the Results section, an analysis of predictive accuracy was conducted for regions independently, to supplement the aggregated findings that were detailed in Figures 3-4 and Tables 2-3. Regional differences in model performance were observed; the median and 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.

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 were quite small. 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.

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.

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 calculations using AMI data. These results can be used to build confidence in model robustness, and also to pre-vet M&V plans for specific projects, given project requirements for uncertainty in reported savings. 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 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.

Future work will focus on four key areas: 1) demonstration of these automated approaches in partnership with utilities, using data from buildings that have participated in whole-building 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) investigation of how these approaches that use measured pre-measure energy use data as the baseline from which savings are calculated, can comport with evaluation requirements to consider code baselines.

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 the members of the project Technical Advisory Committee for their participation and feedback throughout the course of the work. We also acknowledge each of the developers who submitted baseline models for inclusion in this study, and those who contributed to the test dataset.

References

ASHRAE. ASHRAE Guideline 14-2002, Measurement of Energy and Demand Savings. American Society of Heating Refrigeration and Air Conditioning Engineers, ISSN 1049-894X, 2002.

Efficiency Valuation Organization (EVO) (2012). International performance measurement and verification protocol: concepts and options for determining energy and water savings, 1, EVO 10000-1:2012.

Granderson, J, Price PN, Jump, D, Addy N, Sohn, MD. 2015. Automated measurement and verification: Performance of public domain, whole-building electric baseline models. Applied Energy 144: 106-113.

Granderson, J, Price, P, Jump, D, Sohn, M. 2014. Development and application of a statistical methodology to evaluate the predictive accuracy of building energy baseline models. Energy 66(1): 981-990.

Haberl JS, Thamilseran, S. 1998. The great energy predictor shootout II: Measuring retrofit savings. ASHRAE Journal, 40(1):49-56.

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.

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, 2013. Report Number E13-264.

Kreider, JF, Haberl, JS. 1994. Predicting hourly building energy use: The great energy predictor shootout — Overview and discussion of results. ASHRAE Transactions, 100(2):1104-1118.

Mills, E. 2011. Building commissioning: A golden opportunity for reducing energy costs and greenhouse gas emissions in the United States. Energy Efficiency 4(2):145-173.

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, December 2013. LBNL-6611E.