Performance Metrics and Objective Testing Methods for Energy Baseline Modeling Software

Jessica Granderson, Phillip N. Price and Michael D. Sohn
Lawrence Berkeley National Laboratory
David Jump, Quantum Energy Services and Technologies

ABSTRACT

With advances in energy metering, communication, and analytic software technologies, providers of Energy Management and Information Systems (EMIS) are opening new frontiers in building energy efficiency. Through their engagement platforms and interfaces, EMIS products can enable energy savings through multiple strategies including equipment operational improvements and upgrades, and occupant behavioral changes. These products often quantify whole-building savings relative to a baseline period using methods that predict energy consumption from key parameters such as ambient weather conditions and operation schedule. These automated baseline models streamline the M&V process and are of critical importance to owners and utility program stakeholders implementing multi-measure energy efficiency programs.

This paper presents the results of a PG&E Emerging Technology program, undertaken to advance capabilities in evaluating EMIS products for building-level baseline energy modeling. A general methodology to evaluate baseline model performance was developed and used with hourly whole-building energy data from nearly 400 small and large commercial buildings. Evaluation metrics describing model accuracy were identified and assessed for their appropriateness in describing model baseline performance, as well as their usefulness for identifying and pre-screening buildings for whole-building savings estimation suitability. The state of five public-domain models was assessed using the methodology and test data set, and implications for whole building M&V described. Finally a protocol was developed to test EMIS vendor's proprietary models while navigating practical issues concerning test data security, vendor intellectual property, and maintaining appropriate testing 'blinds,' while processing a large data set. Ongoing work entails stakeholder vetting, demonstration of the test procedures with new baseline models solicited from the public, and publication of the results for industry adoption.

Introduction

Energy Management and Information Systems (EMIS) span a spectrum of technologies and services including energy information systems (EIS), building automation systems, fault detection and diagnostics, and monthly energy analysis tools. Tools such as EIS have enabled whole-building energy savings of up to 10-20% with simple paybacks on the order of 1-3 years (Granderson 2009, 2013) through multiple strategies such as: identification of operational efficiency improvement opportunities, fault and energy anomaly detection, and inducement of behavioral change among occupants and operations personnel.

In addition to *enabling* operational savings, some EMIS offerings also automate the *quantification* of whole-building energy savings, relative to a baseline period, using empirical

baseline models that relate energy consumption to key influencing parameters, such as ambient weather conditions and building operation schedule (Granderson 2011; Kramer 2013a, 2013b; Reddy 1997). Today, the advent of increasingly available interval meter data has enabled the development of more robust baseline models than the monthly models that have traditionally been used to characterize whole-building energy performance (Haves 2014; Katipamula 1998; Walter 2103). These automated baseline models can be used to streamline the whole-building measurement and verification (M&V) process. This is important because traditional M&V processes using engineering calculations can comprise a significant portion of the total costs of efficiency programs, and require a level of engineering expertise that can challenge scalability.

Although EMIS hold great promise, several questions remain to be answered before energy managers and utility programs can confidently adopt their emerging M&V automation capabilities (Kramer 2013b). This paper documents research findings that begin to address some of these questions, namely:

1. *How can baseline models be objectively evaluated* to determine general performance robustness for M&V of energy efficiency savings?
2. *What is the state of public domain models*, i.e., what is their accuracy, and what are the associated implications for automated whole-building measurement and verification?
3. *How can public and proprietary software tools be tested and compared*, i.e. what are the elements of a testing protocol, and what blinds must be incorporated into the process.

Extending prior research (Granderson 2012, Granderson 2014), we present a statistical methodology to evaluate the predictive accuracy of baseline energy models used for automated whole-building savings quantification, and apply the methodology to assess the performance of industry-standard models commonly used by M&V professionals, and commercial EMIS offerings. Lengthy periods of interval meter data from several hundreds of buildings are collated to form a 'test' data set, and statistical cross-validation is performed to gauge performance relative to the M&V-focused metrics and time scales of interest. This methodology shares important similarities to the approaches used in the ASHRAE 'shootouts' of the mid and late 1990s (Haberl 1998; Kreider 1994). In both cases, cross-validation is used to determine model error, and in both cases, normalized root mean squared error is included as a performance metric. However, the ASHRAE shootouts were limited to data from a total of two buildings, and the cross-validation was conducted only for short subsets of the model training period.

An important feature of the present work is that the methodology can be used to objectively assess the predictive accuracy of a model, without needing to know the specific algorithm, or underlying form of the model. Therefore, proprietary tools can be evaluated while protecting the developer's commercial intellectual property. The findings of this work can be used to (1) inform technology assessments for EMIS and other technologies that deliver operational and/or behavioral savings; and (2) set a floor of performance of automated M&V, that can be used to set requirements for efficiency programs, including the tradeoffs between cost, and accuracy.

Baseline Model Performance Assessment Methodology

Baseline energy use models characterize building load or consumption according to key explanatory variables such as time of day, and weather. These baseline models are used for a variety of purposes in EMIS, including near real-time energy anomaly detection, and near future

load forecasting, as well as quantification of energy or demand savings (Granderson 2009, 2011). Baseline model accuracy is critical to the accuracy of energy savings that are calculated according to the IPMVP. For both whole-building and measure isolation approaches (IPMVP Options B and C) the baseline model is created during the "pre-measure" period, before an efficiency improvement is made. The baseline model is then projected into the "post-measure" period, and energy savings are calculated based on the difference between the projected baseline and the actual metered use during the post-measure period (EVO 2012). Therefore, the error in reported savings is proportional to the error in the baseline model forecasts.

General Methodology

Prior work established a general 4-step statistical procedure that can be used to evaluate the performance, i.e. predictive accuracy, of a given baseline model (Granderson 2012). This process is described below, and illustrated in Figure 1.

1. Gather a large test data set comprised of interval data from hundreds of commercial buildings.
2. Split the test data from each building into two time periods, the "training" period and the "prediction" period. These periods can be chosen according to the specific application, or use case of interest, e.g., for quantifying energy efficiency savings there is a need to predict baseline energy use over many months, so the timescale of interest is on the order of several months to one year.
3. For a given set of baseline models, generate predictions based on the training period data, compare those predictions to the data from the prediction period, and compute statistical performance metrics based on the comparison. Again, the models of interest, and the specific performance metrics can be tailored to according to the specific application or use case.
4. Assess relative and absolute model performance using the performance metrics that were computed in Step 3.

The accuracy of model predictions for a system or building depends on the robustness of the model, as well as the variability in control, operations, and use of the specific building or system. This testing methodology assesses model performance in general, 'on average' across populations of many buildings; it is not intended to reveal whether a given model will provide accurate results for a *specific* building or project.

Definition of Specific Parameters

Building upon this *general* 4-step process, specific parameters relevant to the whole-building M&V application were defined, as described in the following.

Test data set: Whole-building baseline models can include any number of independent variables that are then used to predict building load or energy use. In the most commonly-used models, outside air temperature, and day/time information from the interval meter time stamp are the only independent variables. Outside air temperature is readily available from building location and weather feeds, whereas models that used other independent variables were not accessible to the research team.

The analyses presented used a multi-year data set of interval meter data that was randomly selected from mid-size commercial customers across a large utility territory. This representative dataset included electricity data from about 400 buildings. We found that sample size was large enough to estimate the statistical distribution of baseline model errors for mid-sized commercial buildings *as a whole* – in fact, even a random sample of 30 or 40 buildings would have been adequate.

Training and prediction periods: Given the whole-building M&V application case, a twelve-month prediction period was deemed of most interest by external stakeholders. This is due to the fact that one year is the typical time period used to quantify efficiency project savings and payouts, and the fact that one year pre- and post-measure data are recommended in ASHRAE Guideline 14 (ASHRAE 2002). Given a desire to shorten the overall M&V process, and therefore total project time, we also considered three, six, and twelve-month training periods in evaluating model performance.

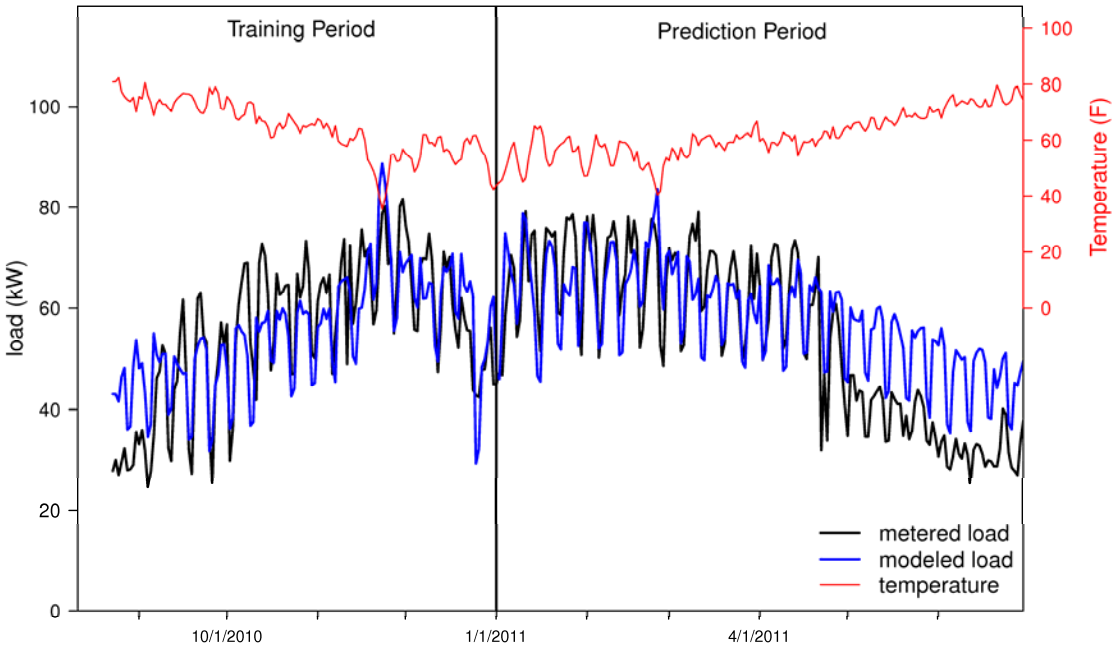

Figure 1. Illustration of steps two and three in the general methodology to evaluate baseline model performance.

Performance metrics: For whole-building measurement and verification (M&V) of energy efficiency measures, a key metric of performance is the error in the total amount of energy used during an evaluation period. The error in total energy use during the prediction, or post-measure period, is referred to as the *bias*. The absolute percent bias error, APBE, is the metric used in this work to quantify error in the total energy use predicted by the model. It is defined in Equation 1 where E_{total} is the measured energy use, $\hat{E}_{total}$ is the model predicted energy use and N is the total number of measurements.

$$APBE = \left| \frac{\hat{E}_{total} - E_{total}}{E_{total}} \times 100 \right| \text{ or } APBE = \left| \frac{\sum_{i=1}^{N} \hat{E}_i - \sum_{i=1}^{N} E_i}{\sum_{i=1}^{N} E_i} \times 100 \right| \quad (1)$$

The second performance metric of interest relates to the ability to predict the total energy used for each individual month. This ability is desirable because if a model fits individual months well then it may be possible to reduce the duration of either the baseline period or the evaluation period. Additionally, if a model generally predicts well for individual months, but a few months stand out as being poorly predicted, this can help to locate problems that need attention and that might affect the efficacy or assessment of the energy efficiency measure. The *Mean Absolute Percent Error (MAPE)* in the monthly energy predictions is defined in Equation 2. The MAPE metric is conceptually very similar to the coefficient of variation of the root-mean-squared error CV(RMSE), which is used in ASHRAE Guideline 14, which is a more common metric in the industry. Monthly MAPE and CV(RMSE) were both investigated; we found that monthly MAPE proved marginally more useful for discriminating between buildings that have less- or more-predictable energy use.

$$MAPE_{month} = \frac{\sum_{m=1}^{12} 100 \times \left| \frac{\hat{E}_m - E_m}{E_m} \right|}{12} \quad (2)$$

Baseline models: Five 'open-source' models from the public domain literature were evaluated. They include change point models, monthly degree-day models, and hourly regression models, and are detailed in the Appendix. These models were selected because they were readily accessible, and representative of the current state of common engineering practice, and EMIS technologies - *not* because they are unique, or were deemed to be the *best* whole-building baseline models. They were used as reference cases to establish a 'benchmark', or 'floor' for the accuracy of automated M&V. This performance benchmark can be used to interpret the performance of baseline models used in proprietary tools – one would not logically elect to use a tool that fares worse than published open source methods.

Results

State of Public Domain Models

Table 1 summarizes the percentiles and mean absolute percent bias error (APBE) for each model, using 12-month training and prediction periods. The mean APBE for the public domain models was approximately 8.4%, and for half of the buildings in the data set, it was less than 5%. This suggests that for large representative samples and one-year pre- and post- M&V conditions, models that exhibit mean APBE much greater than 8% or median biases much greater than 5% would not measure up to the public domain models that are currently available, and may not be as appropriate for whole-building M&V *in general*. Of course, those models may

exhibit much better performance for specific, well-behaved *individual buildings*, with highly predictable loads.

For the monthly MAPE metric, mean monthly MAPE for the public domain models ranged from approximately 16% to 21%, as summarized in Table 2. For half of the buildings in the data set, monthly MAPE was often less than 10%. This suggests that for large representative samples and one-year pre- and post- M&V conditions, models that exhibit mean MAPE much greater than 20% or median MAPE much greater than 10% would not measure up to the currently available public domain models.

Table 1. Percentiles and mean of absolute percent bias error for the 389 buildings in the representative data set, for each model; 12-month training period, 12-month prediction period

Model	10%	25%	50%	75%	90%	Mean
Mean Week	0.82	2.21	4.82	9.63	19.42	8.40
Monthly CDD and HDD	0.69	2.09	4.53	10.03	19.38	8.46
Day, Time, and Temperature	0.69	2.17	4.51	9.26	19.41	8.42
Day and Change Point	0.73	2.02	4.70	9.22	18.84	8.24
Time of Week and Temperature	0.82	2.21	4.82	9.63	19.42	8.40

Table 2. Percentiles and mean of monthly mean absolute percent error for the 389 buildings in the representative data set, for each model; 12-month training period, 12-month prediction period

Model	10%	25%	50%	75%	90%	Mean
Mean Week	5.72	8.80	13.80	23.10	38.30	21.51
Monthly CDD and HDD	4.10	5.40	8.80	16.30	32.64	16.39
Day, Time, and Temperature	3.19	5.00	8.30	15.57	31.20	15.88
Day and Change Point	4.22	6.30	10.2	17.90	33.58	17.50
Time of Week and Temperature	3.20	4.90	8.10	15.50	31.16	15.76

Relative Model Performance

When considering a 12-month training period and 12-month prediction period there was relatively little difference in performance between the five public domain models. The median absolute percent bias is between 4.5 - 4.8% for all of the models, and the mean is between 8.3 - 8.5% (see Table 1). There are a few buildings for which the predictions are extremely poor, with errors greater than 75% (in either direction), and these led to the average being much worse than the median. For the monthly MAPE metric, the range in relative performance was slightly larger than for bias: the medians for the various models range from about 8 - 14%, and the means range from about 16 - 22%. Depending on the specific data set and buildings used, the values achieved for a given performance metric will differ. The results reported here correspond to a random sample of buildings from a large utility territory. When the training period was reduced to 6 months, there was not a significant degradation in median error relative to cases in which 12 months of training data were provided. The exception was the monthly CDD and HDD model, which performed worse on average than models that used interval data. When the training period was reduced even farther, to only 3 months, errors rose significantly, and the time-of-week-and-

temperature and day-time-and-temperature models consistently outperformed the others. Although whole-building M&V guidelines tend to focus on 12-month training and prediction periods, there is a desire to shorten the time required for M&V, which motivated the investigation of shorter training periods; shorter prediction periods may be used for normalized as opposed to avoided energy savings calculations and are also of interest.

Portfolio Aggregation Effects

The results discussed so far have focused on distributions of errors for collections of many individual buildings. However, prediction errors are much smaller when aggregated over a collection of buildings that are treated as a group. A portfolio of buildings will usually include some in which the prediction is too low and others in which it is too high. Although the magnitude of the error will tend to increase as buildings are added to a portfolio, the relative error will tend to decrease or remain stable/constant.

The reduction of errors due to aggregating buildings into a portfolio was explored by grouping buildings with similar uses, based on knowledge of the NAICS code for each building in the test data set. For example, retail stores and public administration buildings might form separate portfolios. In all of these cases the percent bias in the prediction of the portfolio's energy use is less than the mean bias for the individual buildings of that type, because of the aggregation effects discussed above. Table 3 shows the aggregation of buildings by NAICS code, and that the percent bias for the portfolio is often less than 2%. In contrast, without aggregation, the median percent biases by NAICS code was found to range from 2.7 to 7.3.

Table 3. Percent Bias Error for portfolios based on NAICS code, for the time-of-week-and-temperature model

NAICS code	Bldgs	Total kWh	Predicted kWh	Percent bias
42 wholesale trade	14	7,844,788	7,696,758	-1.89
44 retail trade	41	29,935,698	30,370,868	1.45
45 retail trade	12	7,320,698	7,358,519	0.52
49 transp./warehousing	10	5,720,874	5,591,634	-2.26
51 information	15	13,770,148	13,601,572	-1.22
53 real est. rental/leasing	53	37,462,843	41,062,271	9.61
61 educational services	42	16,88,7745	17,403,489	3.05
62 health care/soc. assist.	36	20,238,549	21,001,653	3.77
71 arts/entertainment/rec	30	7,430,195	7,573,492	1.93
72 accomod./food services	63	23,302,962	22,971,386	-1.42
81 other services	32	7,303,410	7,447,883	1.98
92 public administration	6	5,127,729	5,215,852	1.72

Software Testing Protocols

The evaluation methodology, developed and tested with public domain models, provided the basis for a set of software testing protocols, that account for the fact that baseline models are often embedded into software packages. Two protocols were written: 1) a prequalifying test protocol in which baseline modeling software predictive accuracy may be evaluated for a target

population of buildings, and 2) a field test protocol, in which accuracy may be evaluated for a particular building. These protocols provide flexibility in the evaluation of software performance depending on the requirements of the interested parties, and are intended as a starting point for further development. Several practical requirements for implementing baseline model performance evaluations are addressed by the protocols, including: building a test data set for the target building population; protecting intellectual property of vendor's proprietary models and software; maintaining data privacy and security; and assuring software test integrity.

The protocols address the vendor intellectual property issue by providing two pathways to conduct the evaluation – by the vendor providing the test administrator with compiled software, or by the test administrator providing the vendor with building data sets to run on their software. The protocols prohibit access to building owner information and describe how data security may be maintained through the application of 'masks.' Appropriate data 'blinds' are described so that prediction period energy use is not shared with vendors. This focuses the evaluation on the quality of model predictions and prevents intervening with software predictions. These protocols will be exercised in a product test demonstration with voluntary participation from software vendors. This demonstration will provide useful insight about useful testing strategies as well as feedback on the performance of selected proprietary models.

Discussion and Conclusions

This work has demonstrated a general statistical methodology to evaluate both public and proprietary baseline model performance. The specific parameters in the general methodology were defined for use in applications focused on whole-building measurement and verification for efficiency programs. Namely, considerations for building up a test data set, performance metrics most relevant to M&V for whole-building energy savings, and training and prediction periods of key interest. This work complements and extends prior research efforts such as the ASHRAE Shootouts of the 1990s (Haberl 1998; Kreider 1994) and a more recent study conducted by Lawrence Berkeley National Laboratory (Granderson 2012).

State of Public Domain Models, and Implications for M&V

This work showed that for a 12-month post-measure installation period, use of a six-month baseline period, i.e., six months of training data, may generate results that are just as accurate as those based on a 12-month baseline period. This has important implications, as reducing the total length of time required for M&V is key to scaling the deployment of efficiency projects in general, and reducing overall costs. Although existing M&V guidelines recommend a full 12 months of pre- and post- data, these guidelines were developed when monthly data was the standard. Improved baseline models that take advantage of increasingly available interval meter data may not require a full 12-months to develop an accurate baseline.

The analyses conducted for this study were useful in illustrating the bounds of performance accuracy that can be achieved when conducting *fully automated* whole-building measurement and verification. That is, the best performance that can be achieved without the oversight of an engineer to identify non-routine adjustments or incorporate knowledge regarding changes in building occupancy or operations. With the public domain models that were available for investigations, and the representative dataset of hundreds of buildings, this work showed median model errors of under 5% and mean errors of less than 9%. When prescreening was conducted to intentionally target participants to minimize baseline errors, the median error

actually increased slightly but the mean error was reduced to under 7%, and most of the least predictable buildings were eliminated, for a screening criterion that was satisfied by half of the buildings. Using a more restrictive screening criterion, even more of the very poorly predictable buildings were eliminated; for the best-performing model, that criterion was satisfied by about a quarter of the buildings and the mean error was reduced to under 6.5%, with 90% of the building baselines being predicted to within 10%.

As typically practiced, M&V is not fully automated, but is conducted *by an engineer* who has access to information about building occupancy, internal loads, and operations. They can therefore apply their expertise and insights to develop baseline 'adjustments' which tailor savings calculations to the particular building being evaluated. For example, in this study 20% of the buildings in a representative sample exhibited large changes in load that might be straightforward for an engineer to identify and account for, but are not easily handled in the fully-automated case. Collectively, these results suggest that modern tools, with their automated baseline models and savings calculations can *at a minimum*, provide significant value in streamlining the M&V process, providing results that could be quickly reviewed by an engineer to determine if further adjustments are necessary. They also suggest that savings can be reliably quantified at the whole-building level, using today's interval data-based models. Depending on the level of confidence required, and the precise depth of savings expected, these savings might be quantified in a fully automated manner, or with some engineering intervention.

Whole-building approaches to savings can include multi-measure savings strategies, including major system and equipment efficiency upgrades, operational improvements, and behavioral programs. This multi-measure approach is expected to yield a higher depth of savings, of up to 20% or more. As a point of reference, retro-commissioning (RCx) alone, saves on average 16% in commercial buildings (Mills 2009). This work showed that a small sample of public domain models demonstrates prediction accuracy within twenty percentage points for 90% of the cases, and within five percentage points for 50% of the cases. With very simple prescreening, accuracy improves by 1-2 percentage points. Note that no such accuracy prediction is available for engineering calculations, which are typically provided for single-measures that amount to 1 to 10% of whole-building energy use. Whole-building savings estimation should therefore be no more risky than engineering calculations.

When buildings are aggregated into a portfolio, errors tend to cancel out so that the percent error in the predicted energy use decreases substantially. Depending on the method of creating the portfolio (e.g. at random, or by screening on the goodness of fit during the training period, or by selecting buildings of a given business type), the total annual energy use of a portfolio of about 40 buildings can usually be predicted within 1.5 – 4% accuracy. The benefits of portfolio aggregation would not impact any individual customer or program participant, but *are* relevant from the perspective of the utility, which may report savings at the aggregated level of many programs, or many buildings. This also has implications for improved confidence related to regulatory and evaluation considerations, and increased ability to realize deeper savings from multi-measure whole-building focused efficiency programs.

Future Work

The analyses in this study made use of freely available public domain reference models to determine the general state of some of the whole-building baseline models that are commonly used by today's engineers. This study did not focus on identifying the *best* whole-building baseline models, an exercise that would ideally include a diversity of proprietary as well as

public models. (Jump 2013) began to establish protocols that integrate the model evaluation methodology with the blinds necessary to protect data privacy and the intellectual property underlying proprietary baseline models; applying these protocols to the testing of commercial tools to validate scalability and practicality is a key next step.

While this paper focused whole-building applications, the assessment methodology is general, and therefore can also be used to evaluate baseline models for applications such as continuous energy anomaly detection, demand savings, or system-level isolation approaches to M&V. Future work will therefore also focus on defining the most appropriate performance metrics, time horizons, and test data sets for this extended set of use cases for baseline models, beginning with isolation-based approaches. An important next step involves extensive industry engagement to build conceptual awareness and buy-in, and technical vetting. A call will be issued to solicit novel and unique baseline models from the public, and these models will be tested with the methodology from this paper, evaluated, and published in the public domain to facilitate more widespread adoption of M&V methods that promise streamlining through automation. Finally, this study focused on the *general* assessment of M&V baseline model performance accuracy across large populations of buildings. It did not delve into the most rigorous means of quantifying the uncertainty in reported savings once an actual project has been conducted. Such a study would also set the stage to compare the uncertainty in reported savings that results from the use of measured approaches, versus those that result from the use of engineering calculations.

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 work described in this paper was also funded by the Pacific Gas and Electric Company, and was developed as part of Pacific Gas and Electric Company's Emerging Technology program under internal project number ET12PGE1311. The authors also acknowledge all others who assisted this project, including: PG&E's Leo Carillo, Mananya Chansanchai, Mangesh Basarkar, and Ken Gillespie; Cody Taylor of DOE's Building Technologies Office; Portland Energy Conservation Inc., Agami Reddy, and Technical Advisory Group members.

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). International Performance Measurement and Verification Protocol: Concepts and options for determining energy and water savings, Volume I. January 2012. EVO 10000-1:2012.

Granderson, J, Piette, MA, Ghatikar, G, Price, PN. Building energy information systems: State of the technology and user case studies. Lawrence Berkeley National Laboratory, November 2009, LBNL-2899E.

Granderson, J, Piette, MA, Rosenblum, B, Hu, L, et al. Energy information handbook: Applications for energy-efficient building operations. Lawrence Berkeley National Laboratory, 2011, LBNL-5272E.

Granderson J, Price PN. Evaluation of the Predictive Accuracy of Five Whole-Building Baseline Models, Lawrence Berkeley National Laboratory, 2012, LBNL-5886E.

Granderson, J, Lin, G, Piette MA. Energy information systems (EIS): Technology costs, benefits, and best practice uses. Lawrence Berkeley National Laboratory, 2013, LBNL-6476E.

Granderson, J, Price PN. 2014. Development and application of a statistical methodology to evaluate the predictive accuracy of building energy baseline models. Energy. Published online February 21, 2014.

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

Haberl, J, Culp C, Claridge, D. ASHRAE's Guideline 14-2002 for measurement of energy and demand savings: How to Determine what was really saved by the retrofit. Proceedings of the 5th International Conference for Enhanced Building Operations, October 2005.

Haves, P, Wray, C, Jump, D, Veronica D, Farley, C. Development of diagnostic and measurement and verification tools for commercial buildings**.** Report prepared for California Energy Commission. 2014.

Jump, D, Price, PN, Granderson J, Sohn MD. Functional testing protocols for commercial building efficiency baseline modeling software. Pacific Gas and Electric, 2013, ET Project Number ET12PGE5312.

Katipamula, S, Reddy, TA, Claridge, DE. 1998. Multivariate regression modeling. Journal of Solar Energy Engineering, Transactions of the ASME 120(3):177-184.

Kramer, H, Effinger, J, Crowe E. Energy management and information system (EMIS) software technology assessment: Considerations for evaluating baselining and savings estimation functionality. Pacific Gas and Electric, 2013, ET Project Number ET12PGE1311.

Kramer, H, Russell, J, Crowe, E, Effinger, J. Inventory of Commercial Energy Management and Information Systems (EMIS) for M&V Applications, Northwest Energy Efficiency Alliance, 2013, #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.

Mathieu, JL, Price, PN, Kiliccote, S, Piette, MA. Quantifying changes in building electricity use, with application to Demand Response. IEEE Transactions on Smart Grid 2:507-518, 2011.

Reddy, TA, Saman, NF, Claridge, DE, Haberl, JS, Turner WD, Chalifoux AT. 1997. Baselining methodology for facility-level monthly energy use – Part 1: Theoretical aspects. AHSRAE Transactions 103(2):336-347.

Walter T, Price PN, Sohn MD. Uncertainty estimation improves energy measurement and verification procedures. Submitted to Applied Energy, September 2013.

Appendix

This appendix details the five whole-building baseline models that were included in this study. In the M*ean-Week (MW)* model, the 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.

The *Cooling-Degree-Day and Heating-Degree-Day (CDD-HDD)* model represents techniques that were originally developed to analyze monthly utility billing data. For each month linear regression is performed to predict monthly energy usage as a function of CDD and HDD, using base temperatures of 55 F and 65 F, respectively. With m identifying the month, the model can be expressed according to Equation A-1 as:

$$E_m = \beta_0 + \beta_C CDD_m + \beta_H HDD_m \quad \text{(A-1)}$$

The *Change-Point* model implements a six-parameter change-point model with the addition of a day-of-the-week effect. Detailed in (ASHRAE 2002; Haberl 2005), 5-parameter change-point models include: the slope of the load-vs-temperature line for low temperatures, the slope of the line for high temperatures, the change point below which the temperature is low, the change point above which it is high, and the average load for temperatures that are neither low nor high. In this study, there were enough data to estimate and implement more parameters: (1) estimated slope for intermediate temperatures, and (2) at the suggestion of a subject matter expert, the change-point model also allowed each day of the week to have a different average load in the intermediate-temperature region.

In the *Day-Time-Temperature* model the predicted load is a sum of several terms: (1) a "day effect" that allows each day of the week to have a different predicted load; (2) an "hour effect" that allows each hour of the day to have a different predicted load; (3) an effect of temperature that is 0 for temperatures above 50F and is linear in temperature for temperatures below 50F; and (4) an effect of temperature that is 0 for temperatures below 65F and is linear in temperature for temperatures above 65F. We define the following: i identifies the data point, day_i and $hour_i$ are the day and hour of that data point; $TC_i = 0$ if the temperature T exceeds 50 and is equal to 50 F -T if T < 50 F; TH=0 if T < 65 F and is equal to T-65 F if T > 65 F. With these definitions, the *Day-Time-Temperature* model can be written as:

$$E_i = \beta_{dayi} + \beta_{houri} + \beta_C T_{Ci} + \beta_H T_{Hi} \quad \text{(A-2)}$$

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

Mathieu et al. (2011), but the determination of "occupied" and "unoccupied" periods is new to this project. For each day of the week, the 10th and 90th percentile of the load were calculated; call these L10 and L90. The first time of that day at which the load usually exceeds the L10 + 0.1*(L90-L10) 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.