Applied Energy 144 (2015) 106–113

Contents lists available at ScienceDirect

Applied Energy

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

Automated measurement and verification: Performance of public domain whole-building electric baseline models

Jessica Granderson [a,*], Phillip N. Price [a], David Jump [b], Nathan Addy [a], Michael D. Sohn [a]

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

HIGHLIGHTS

- All five tested baseline models predict annual energy use with similar accuracy.
- A twelve-month training period leads to errors less that 8.5% in most buildings.
- A shorter training period is sufficient for models that adjust for weather.
- Combining buildings into a portfolio leads to lower relative error in energy use.
- Evaluation methods can be applied to black-box models.

ARTICLE INFO

Article history:
Received 21 January 2014
Received in revised form 20 November 2014
Accepted 7 January 2015

Keywords:
Baseline prediction
Energy savings
Performance accuracy
Whole-building energy
Energy efficiency programs
Energy management and information systems

ABSTRACT

We present a methodology to evaluate the accuracy of baseline energy predictions. To evaluate the predictions from a computer program, the program is provided with electric load data, and additional data such as outdoor air temperature, from a "training period" of at least several months duration, and used to predict the energy use as a function of time during the subsequent "prediction period." The predicted energy use is compared to the actual energy use, and errors are summarized with several metrics, including bias and mean absolute percent error (MAPE). An important feature of this methodology is that it can be used to assess the predictive accuracy of a model even if the model itself is not provided to the evaluator, so that proprietary tools can be evaluated while protecting the developer's intellectual property. The methodology was applied to evaluate several standard statistical models using data from four hundred randomly selected commercial buildings in a large utility territory in Northern California; the result is a statistical distribution of errors for each of the models. We also demonstrate how the methodology can be used to assess the uncertainty in baseline energy predictions for a portfolio of buildings, which is an issue that is important for the design of utility programs that incentivize energy savings. The findings of this work can be used to (1) inform technology assessments for technologies that deliver operational and/or behavioral savings; and (2) determine the expected accuracy of statistical models used for automated measurement and verification (M&V) of energy savings.

1. 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 [1]. 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 [2,3] 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* savings, some EMIS also automate the quantification of whole-building energy savings, relative to a baseline period, using empirical models that relate energy consumption to parameters such as ambient weather conditions and building operation schedule [4–8]. Interval meter data enable the use of baseline models that have several advantages over the monthly models that have traditionally been used to characterize whole-building energy performance [9–11]: they can determine the

* Corresponding author at: 1 Cyclotron Rd., MS 90-3111, Berkeley, CA 94720, USA. Tel.: +1 (510) 486 6792.
E-mail address: JGranderson@lbl.gov (J. Granderson).

http://dx.doi.org/10.1016/j.apenergy.2015.01.026
0306-2619/

relationship between temperature and electric load more accurately and with a shorter duration of data, and they can make predictions at a much finer timescale.

Automated baseline models can be used to streamline the whole-building measurement and verification (M&V) process, greatly reducing the cost compared to traditional processes, which require a level of building engineering expertise that limits scalability. However, several questions remain to be answered before energy managers and utility programs can confidently adopt emerging automation capabilities. For example, in energy efficiency applications one objective is to quantify and minimize the uncertainty in reported whole-building savings, which depends on baseline model effectiveness, building predictability, portfolio aggregation effects, and depth of savings being measured [6].

This paper presents an extension of prior research [5] on how to assess the accuracy and usefulness of whole-building energy models by testing predictions of baseline energy use against actual energy use. We demonstrate the method by applying it to a large random sample of commercial building data to answer the following questions of practical importance:

1. What is the *state of public domain models*, i.e., how well do they perform, and what are the associated implications for automated whole-building measurement and verification?
2. How can buildings be pre-screened to identify those that are highly model-predictable and those that are not, in order to identify estimates of building-energy savings that have small errors/uncertainty?

In this study we evaluated only public-domain whole-building baseline energy models for which outdoor air temperature is the only predictive variable. However, the methodology that we used can be applied to evaluate any model, including models that make use of additional data such as occupancy levels, business types, or building types.

While resources such as ASHRAE Guideline 14 and the International Performance Measurement and Verification Protocol (IPMVP) [12,13], 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. They also provide little guidance on using interval data as opposed to monthly data. The methodology presented in this work extends the principles in these existing resources to quantify model predictive accuracy after the training period, and suggests key performance metrics to quantify model accuracy in the context of whole-building M&V. 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 [14,15]. 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 this 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 products 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 consider requirements for utility or corporate efficiency programs, including the tradeoffs between cost, and accuracy.

2. 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 [2,4].

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 [13]. Therefore, the error in reported savings is proportional to the error in the baseline model forecasts.

2.1. 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 [5].

(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 model training and model prediction periods. These periods can be tailored according to the specific application or use case of interest, e.g., energy efficiency savings, demand response load reductions, or continuous energy anomaly detection. For this study, the focus was measurement and verification of energy savings at the whole-building level.
(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.

This process is illustrated in Fig. 1, which shows daily average loads. The first several months constitute the "training period," from which the load data and outdoor air temperature data are used to create a statistical model that predicts load as a function of the time during the week and the temperature. This model is then used to predict the load during both the prediction period and the subsequent training period.

The subject building for Fig. 1 has several features that are typical of commercial buildings: the load is temperature-dependent, and load on weekends is substantially lower than the load on weekdays. At a finer timescale, the building also has a nightly minimum load that is much lower than the daytime maximum, but of course this cannot be seen on this plot of daily averages. (Plotting 10 months of hourly or 15-min data would create a plot with so many vertical oscillations that it would be impossible to interpret).

Furthermore, sometime around the beginning of May 2011 the building's energy behavior changes: both the weekday and week-

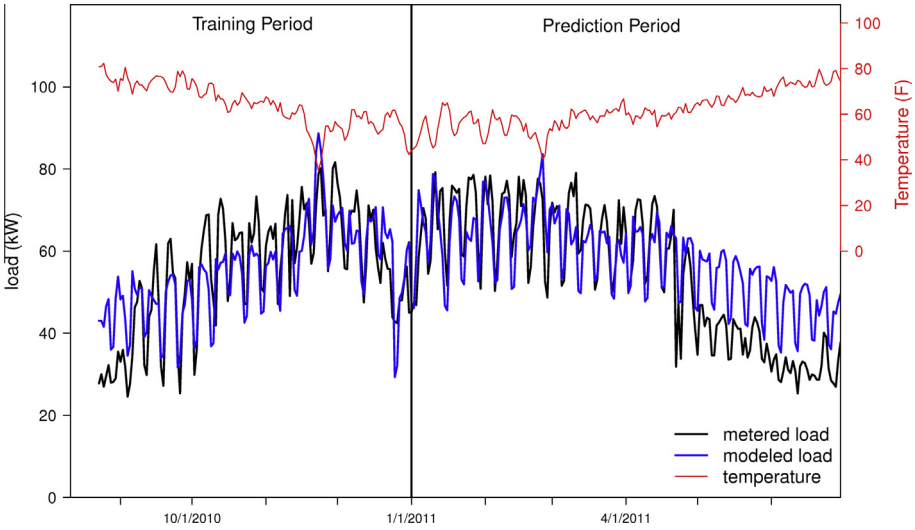

Fig. 1. Illustration of steps two and three in general methodology to evaluate baseline model performance.

end load decrease substantially, with the weekday load decreasing more than the weekends. Sudden or gradual changes of about this magnitude are fairly common in commercial building load data, as we discuss in this report, and are a major cause of the error in baseline model predictions.

The accuracy of model predictions for a system or building depends on how well the functional form of the model captures the effect of the input parameters, as well as the variability in control, operations, and other parameters that are not inputs to the model. 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.

2.2. 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.

2.2.1. 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 used. 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 is large enough to estimate the statistical distribution of baseline model errors for mid-sized commercial buildings *as a whole* – even a random sample of 30 or 40 buildings appears to be adequate. However, we found that it is not large enough to distinguish differences in model performance between *different building types*.

2.2.2. 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 [12]. 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.

2.2.3. 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*, and is defined in Eq. (1) where E_{total} is the measured energy use and $\widehat{E}_{total}$ is the model predicted energy use. A positive bias means the model predicted energy use higher than was measured.

$$B = \widehat{E}_{total} - E_{total} \tag{1}$$

Bias may be evaluated over any time interval, and bias from different sub-intervals can cancel out; for example, if the bias is +7000 kW h in one month, and −7000 kW h in the following month, then the bias for the two-month period will be zero.

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 Eq. (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.

$$\mathrm{MAPE}_{month} = \frac{\sum_{m=1}^{12} 100 \times \left|\frac{\widehat{E}_m - E_m}{E_m}\right|}{12} \tag{2}$$

In addition to Bias, MAPE, and CV(RMSE), many other measures of model accuracy could be used to highlight different aspects of performance. As an anonymous reviewer pointed out, in some applications there is special interest in predicting the daily peak load, and a measure of model fit could be employed that gives high weight to the peak load, or to the hours at which load normally peaks.

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 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.

2.3. Prescreening for higher accuracy

The uncertainty in whole-building savings calculation for a given building is due to the robustness of the baseline model used to determine those savings, as well as the predictability of the building itself. Three potential indicators of model accuracy were investigated to determine their suitability for pre-screening. First a North American Industry Classification System (NAICS) code was considered. Of the several hundred buildings in the test data set, buildings with similar NAICS codes were grouped to determine whether certain building types were predicted with more or less accuracy than others.

Second, a load variability metric was considered that quantifies the degree to which the building's load varies from week to week at the same time during the week, as defined in Eq. (3). For the several hundred buildings in the test data set, the predictive error was compared to the magnitude of the variability metric to determine if there was a correlation. To calculate the metric, the average load is calculated for each time interval during the week (there are 672 15-min time intervals in a week). For each data point, the difference between the load and the average load at that time of the week is calculated and squared. The square root of the average of these squared errors defines the load variability metric, LV.

$$\mathrm{LV} = \sqrt{\frac{\sum_{w=1}^{52}\sum_{t=1}^{672}(y_{w,t} - \bar{y}_t)^2}{(52)(672)}} \tag{3}$$

Finally, goodness of fit during the training period was explored to determine whether buildings which provide a better fit to the energy used during each month of the training period tend to have more predictable total energy use during the prediction period. To quantify the model fit during the training period, we considered both the monthly Mean Absolute Percent Error (MAPE), defined in Eq. (2), and the CV(RMSE) in the predicted monthly energy use.

3. Results

3.1. State of public domain models

Table 1 summarizes the percentiles and mean absolute percent *bias* for each model, using 12-month training and prediction periods. The mean absolute bias for the public domain models was approximately 8.4%, and for half of the buildings in the data set, bias was less than 5%. This suggests that for large representative samples and one-year pre- and post-M&V conditions, models that exhibit mean biases 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.

Table 1
Percentiles and mean of absolute percent bias 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

As shown in Table 2, monthly MAPE for the public domain models ranged from approximately 16% to 21%. 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.

3.2. Pre-screening findings

Of the three screening criteria investigated, only one, monthly MAPE during the training period, proved to be moderately useful. Two screens were evaluated – one in which MAPE during the training period was less than 5%, and one in which MAPE was less than 3%. (Less restrictive screens would have been ineffective; see Price et al. [16] for more detail on this topic). Model performance according to absolute percent bias for the 3% screen are shown in Table 3; results for the change point model were not calculated due to a programming error.

Unexpectedly, this screen did not lead to an overall shift of the distribution toward lower bias – the 10th through 50th percentiles are actually worse than in the dataset without the screening – but it did help substantially at the bad end of the distribution by eliminating many of the worst-fitting buildings. Before applying this screen, even a 20% reduction in energy use would not be detected in 10% of the buildings in the full dataset (see the '90%' column of Table 1), but could easily be detected in all or almost all of the buildings in the screened dataset (Table 3). The performance at the 90th percentile is greatly improved, but the screening only moderately improves the mean absolute bias, from above 8% to less than 7%: a few buildings with absolute percent bias exceeding 30% are still included, which pulls up the mean.

Applying the screening criterion reduces errors, but also reduces the number of 'eligible' buildings. For all but one model, more than half of the buildings met the screening criterion of monthly MAPE < 5%, and the size of the dataset was kept to hundreds. However, for monthly MAPE < 3%, the effect was larger. This

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

Table 3
Percentiles of absolute percent bias for each model, for buildings where monthly MAPE in the training period was less than 3%. *N* shows the number of buildings (out of 389 total) that pass the screen; *N* differs from model to model and is much lower for the mean week model.

Model	*N*	10%	25%	50%	75%	90%	Mean
Mean week	23	3.48	4.10	5.20	5.90	8.32	6.47
Monthly CDD and HDD	72	3.40	4.10	5.45	7.43	9.99	6.82
Day, time, and temperature	112	2.70	3.35	4.70	7.55	10.20	6.67
Time of week and temperature	110	2.69	3.32	4.55	7.20	10.10	6.33

reducing effect was largest in the case of the mean week model, where the monthly MAPE < 5% criterion was met for only 62 of the original 398 buildings in the data set, and the <3% criterion was met for only 23. However, for the other models, more than half of the buildings in the dataset met the screening criterion.

3.3. Relative model performance

To evaluate model performance, each model was fit using data from the training period, and the bias and monthly MAPE were evaluated for the prediction period. When considering a 12-month training period and 12-month prediction period, there was little difference in performance between the five public domain models. The median absolute bias is between 4.5% and 4.8% for all of the models, and the mean is between 8.3% and 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), which is why the average is 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% to 14%, and the means range from about 16% to 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.

Perhaps surprisingly, when the training period was reduced to 6 months – February through July – 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. The February through July training period covers most of the outdoor air temperature range experienced in a year by these buildings, and results might be worse if the training period excluded the entire winter or the entire summer. Results might also differ in other climates: all of our buildings came from within a territory a few hundred miles on a side.

As part of our investigation of the results we looked for a systematic relationship between total annual energy consumption and both bias and monthly MAPE (e.g. are higher-consumption buildings likely to be over- or under-predicted) but no such relationship was found.

3.4. 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 which 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.

Table 4
Bias for portfolios based on NAICS code, for the time-of-week-and-temperature model.

NAICS code	Bldgs	Total (kW h)	Predicted (kW h)	Percent bias	Mean Absolute Percent Bias
42	14	7,844,788	7,696,758	−1.89	10.7
44	41	29,935,698	30,370,868	1.45	6.1
45	12	7,320,698	7,358,519	0.52	5.5
49	10	5,720,874	5,591,634	−2.26	8.3
51	15	13,770,148	13,601,572	−1.22	10.6
53	53	37,462,843	41,062,271	9.61	15.1
61	42	16,88,7745	17,403,489	3.05	6.2
62	36	20,238,549	21,001,653	3.77	5.9
71	30	7,430,195	7,573,492	1.93	12.5
72	63	23,302,962	22,971,386	−1.42	5.4
81	32	7,303,410	7,447,883	1.98	10.7
92	6	5,127,729	5,215,852	1.72	4.6

The reduction of errors due to aggregating buildings into a portfolio was explored by grouping buildings with similar uses, based on the NAICS code for each building in the test data set; the two-digit NAICS prefix classifies businesses into broad categories such as retail stores and public administration buildings. The total energy predicted to be consumed in the 12-month prediction period was summed over all of the buildings with a given two-digit NAICS prefix, and compared to the actual energy consumed by the same buildings. 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 4 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 mean absolute percent bias is much higher (shown in the right-most column). For example, the average prediction error for NAICS prefix 42 is nearly 11%, but some of these are over-predictions and others are under-predictions, so that the bias for the portfolio as a whole is only −1.89%.

4. 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 [14,15] and a more recent study conducted by Lawrence Berkeley National Laboratory [5].

The present paper summarizes the most important methods and results that came out of the comparison of five models applied to a large dataset of commercial building data. Additional details, plots, and discussion can be found in an LBNL report about this work [16].

4.1. 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 was 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. Presumably the attention of an engineer would lead to more accurate savings estimates in many buildings, compared to simply directly applying one of the automated baseline models, but such attention is time-consuming and costly, and it may be more efficient to use a fully automated approach for most buildings and refer a building to an engineer for deeper analysis only if there is a good reason to believe the savings estimate for that building might be grossly in error, e.g. if the monthly MAPE in the training period is very high, or if the load change between the pre-ECM and post-ECM period is much larger than expected. Today's fully automated approaches do not yet provide a means of addressing non-routine adjustments, which if neglected can represent a significant source of error, yet can be minimized with an engineer's insight.

Collectively, our 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 adjustments and further tailoring are necessary. They also suggest, that savings can be reliably quantified at the whole-building level, using the interval data-based models that are available today. 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 [17]. This work demonstrated that a small sample of public domain models is able to demonstrate savings accuracy within 20 percentage points for 90% of the cases, and within 5 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 1–10% 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, even though the error for a typical building in the portfolio is much higher than this. Results for aggregation by business type were provided in Table 4; similar findings were found when aggregating 40 buildings selected at random. 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.

4.2. Prescreening to reduce error

Since errors in whole-building measurement and verification are due to the robustness of the baseline model *in combination with* the predictability of the building, it may be possible to determine building- or load-specific characteristics that correlate with smaller errors. These characteristics or metrics might then be used to pre-screen or target program participants. Filtering buildings for which monthly MAPE was less than 5% eliminated many of the most extreme errors for a large dataset.

In most buildings and most years, the largest source of year-to-year change in energy use is neither energy conservation measures nor year-to-year variation in weather, it is changes in characteristics of building operation and occupant behavior such as operating hours, thermostat settings, the number of occupants and the type of activities performed in the building. Surveys of building owners or occupants, or additional types of data (such as occupancy data for hotels and sales data for retail buildings) might allow better ability to screen out unpredictable buildings, as well as better models. For models fit to much less than a full year of data, and therefore to reduce the time required for M&V, the range of temperatures present in the training period has a substantial effect on the accuracy of the model in many buildings. This is an area that should be explored in more detail.

5. Future work

The analyses in this study made use of freely available public domain reference models to determine the general state of the models that are most 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 models, and models that include variables other than outside air temperature, day, and time. Jump et al. [18] 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. Additionally, this study only scratches the surface when it comes to screening to identify predictable buildings, and investigating the effects of shortening the prediction period; both of these subjects invite further research.

This paper focused on application to whole-building measurement and verification, but the baseline model assessment methodology is general: it is also applicable to evaluation of baseline models for continuous energy anomaly detection, for demand savings, or for evaluating system-level models used in a measure isolation approach to M&V. Future work might focus on defining the most appropriate performance metrics, time horizons, and test data sets for this extended set of use cases for baseline models.

This study did not evaluate actual calculations of savings from applying baseline models to data from buildings in which efficiency projects were implemented. Such a study would yield important information regarding the impact on savings uncertainty from (1) duration of pre- and post-measure periods, (2) baseline model deterioration rate (when to re-baseline), (3) post-installation models. At a minimum, that investigation would require extensive load and independent variable data from before and after energy efficiency improvements have been implemented in each building. This would also set the stage for a long-term study to directly compare of the uncertainty in measured approaches to the uncertainty in approaches based on engineering calculations (as opposed to those based on avoided energy use based on projected baseline models of energy use).

Finally, it is worth noting that the approach applied in this study can be used to evaluate the likelihood that a model will be able to identify a given percent savings, at a desired confidence level. While this paper focused on percent bias and MAPE as key metrics of focus, consideration of CV(RMSE) using the same testing methodology can provide important insights into fractional savings uncertainty. Fractional savings uncertainty can be expressed as a function of parameters such CV(RMSE) in the pre/training period, achieved fractional savings, and, the length of the pre and post periods. Resources such as ASHRAE's Guideline 14 provide of "look-up tables" and plots to determine fractional savings uncertainty for particular confidence levels, CV(RMSE values), and measure post periods [12]. Such analyses can be used to understand the percent savings that can a model can reliably detect.

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. This work described in this report was 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 want to acknowledge all others who assisted this project, including: PG&E's Leo Carillo, Mananya Chansanchai, Mangesh Basarkar, and Ken Gillespie; Portland Energy Conservation Inc., Agami Reddy, and the members of our Technical Advisory Group.

Appendix A

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. This is a simplistic 'naïve' model that was intentionally included for comparative purposes.

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 the number of heating and cooling degree-days is calculated, and linear regression is performed to predict monthly energy usage as a function of CDD and HDD. CDD and HDD were defined with base temperatures of 55 F and 65 F, respectively.

With m identifying the month, the model can be expressed according to Eq. (A-1) as:

$$E_m = \beta_0 + \beta_C CDD_m + \beta_H HDD_m \tag{A-1}$$

The *Change-Point* model implements a six-parameter change-point model with the addition of a day-of-the-week effect. Detailed in [12,19], 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. The ASHRAE model assumes that there is no relationship between temperature and load for temperatures that are above the low-temperature region but below the high-temperature region. With months or even a year of data, as in this study, there are enough data to estimate and implement more parameters: (1) we also estimated a slope for intermediate temperatures, generating models with three slopes instead of two, 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 50 F and is linear in temperature for temperatures below 50 F; and (4) an effect of temperature that is 0 for temperatures below 65 F and is linear in temperature for temperatures above 65 F. The use of time-of-day and day-of-week variables is described in [20], in the context of more complicated regression models that include special handling of, e.g., humidity and holidays.

We define the following: i identifies the data point, day_i and hour_i are the day and hour of that data point; $T_{C_i} = 0$ if the temperature T_i exceeds 50 and is equal to 50 F $-T_i$ if $T < 50$ F; $T_{H_i} = 0$ if $T_i < 65$ F and is equal to T_i-65 F if $T_i > 65$ F. With these definitions, the *Day-Time–Temperature* model can be written as:

$$E_i = \beta_{day_i} + \beta_{hour_i} + \beta_C T_{C_i} + \beta_H T_{H_i} \tag{A-2}$$

The model is fit with ordinary regression. It can be thought of a variant of the ASHRAE five-parameter change-point model [12]. Unlike an ASHRAE five-parameter change-point model, it has fixed points for the temperature slopes (at 50 F and 65 F), and it adds time-of-day and day-of-week variation.

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 Mat-

hieu et al. [21], 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.

References

[1] Consortium for Energy Efficiency (CEE). Summary of commercial whole building performance programs: continuous energy improvement and energy management and information systems. Consort Energy Efficiency 2012.

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

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

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

[5] Granderson J, Price PN. Evaluation of the predictive accuracy of five whole-building baseline models. Lawrence Berkeley National Laboratory, LBNL-5886E; 2012.

[6] 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 Electr 2013. ET Project Number ET12PGE1311.

[7] 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, #E13-264; 2013.

[8] Reddy TA, Saman NF, Claridge DE, Haberl JS, Turner WD, Chalifoux AT. Baselining methodology for facility-level monthly energy use – Part 1: Theoretical aspects. AHSRAE Trans 1997;103(2):336–47.

[9] 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.

[10] Katipamula S, Reddy TA, Claridge DE. Multivariate regression modeling. J Sol Energy Eng, Trans ASME 1998;120(3):177–84.

[11] Walter T, Price PN, Sohn MD. Uncertainty estimation improves energy measurement and verification procedures. Appl Energy 2014;130:230–6.

[12] ASHRAE. ASHRAE Guideline 14-2002, measurement of energy and demand savings. American Society of Heating Refrigeration and Air Conditioning Engineers; 2002. ISSN 1049-894X.

[13] Efficiency Valuation Organization (EVO). International performance measurement and verification protocol: concepts and options for determining energy and water savings, vol. I. January 2012. EVO 10000-1; 2012.

[14] Haberl JS, Thamilseran S. The great energy predictor shootout II: measuring retrofit savings. ASHRAE J 1998;40(1):49–56.

[15] Kreider JF, Haberl JS. Predicting hourly building energy use: the great energy predictor shootout — overview and discussion of results. ASHRAE Trans 1994;100(2):1104–18.

[16] Price P, Granderson J, Sohn M, Addy N, Jump D. Commercial building energy baseline modeling software: performance metrics and method testing with open source models and implications for proprietary software testing. Lawrence Berkeley National Laboratory report LBNL-6602E; September 2013.

[17] Mills E. Building commissioning: a golden opportunity for reducing energy costs and greenhouse gas emissions in the United States. Energy Effici 2011;4(2):145–73.

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

[19] 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. In: Proceedings of the 5th international conference for enhanced building operations; October 2005.

[20] Energy and Environmental Economics. Time dependent valuation of energy for developing building efficiency standards. Report prepared for the California Energy Commission; February 2011.

[21] 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.