LBNL-XXXXX

ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

Evaluation of the Predictive Accuracy of Five Whole-Building Baseline Models

J. Granderson and P.N. Price

Environmental Energy Technologies Division

August 2012

Disclaimer

This document was prepared as an account of work sponsored by the United States Government. While this document is believed to contain correct information, neither the United States Government nor any agency thereof, nor The Regents of the University of California, nor any of their employees, makes any warranty, express or implied, or assumes any legal responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by its trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or The Regents of the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof or The Regents of the University of California.

This report documents the evaluation of five building energy baseline models, including a proprietary model from Pulse Energy. This work does not comprise a product endorsement or recommendation by the authors, or by Lawrence Berkeley National Laboratory.

Evaluation of the Predictive Accuracy of Five Whole-Building Baseline Models

Jessica Granderson, Phillip Price
Lawrence Berkeley National Laboratory

Abstract

This report documents the relative and absolute performance of five baseline models used to characterize whole-building energy consumption. The Pulse Adaptive Model[1], multi-parameter change-point, mean-week, day-time-temperature, and LBNL models were evaluated according to a number of statistical 'goodness of fit' metrics, to determine their accuracy in characterizing the energy consumption of a set of 29 buildings. The baseline training period, prediction horizon, and predicted energy quantity (daily, weekly, and monthly energy consumption) were varied, and model predictions were compared to interval meter data to determine the accuracy of each model. Three combinations of baseline training periods and prediction horizons were considered: 6 months of training to generate a 12-month prediction; 9 months of training to generate a 7-month prediction; and 12 months of training to generate a 6-month prediction.

Although there was no single best performer, the study results showed that the LBNL model, the Pulse Adaptive Model, and the day-time-temperature model consistently outperformed the mean-week, and industry standard change-point models. In aggregate, across the three training and prediction periods that were considered, and across the three energy quantities that were predicted, the *median* absolute percent errors for these models ranged from 3-6% of the actual total metered energy consumption. When considering the normalized root *mean* squared error, monthly energy use was predicted with the least error, and daily energy was predicted with the most error. For the LBNL, day-time-temperature, and Pulse Adaptive Model, these errors ranged from 8-18%. As the training period increased and the prediction horizon decreased, the model performance improved a few percent for any given model. Other statistical metrics such as correlation, root mean squared error, and relative bias, were included in the study, in addition to an assessment of the accuracy of Pulse Energy's reported 90% confidence intervals.

The methodology developed and applied for this study addressed baseline accuracy, which comprises the first step in determining the uncertainty in the measurement and verification (M&V) of gross, whole-building energy savings. Future work will expand the analyses to account for uncertainty in the degree to which, excluding all effects but the

[1] As noted in the Disclaimer, this work does not comprise a product endorsement or recommendation by the United States Government or any agency thereof, by The Regents of the University of California, by the Lawrence Berkeley National Laboratory, or by the authors.

EEMs, the building's operation and energy use during the baseline period is equal to that of the post-measure period. This will require data spanning a longer period of time, from buildings where efficiency measures have been implemented. Since the dataset for this study was limited to a small number of buildings, and not fully representative, widely generalizable conclusions cannot be established. Therefore, future work will focus on expansion of the number of datasets, and diversity of buildings to enable more generalizable findings, disaggregation of different building types, and correlation of errors with load variability.

1. Introduction

Whole-building baseline models are used for a variety of energy efficiency applications, and associated analyses. These applications can be automated in software tools such as energy information systems (EIS), and related analytical platforms. Since they characterize the building's 'typical' energy consumption relative to key factors such as weather, these baselines may be used for longitudinal benchmarking, for real-time energy anomaly detection, or for measurement and verification (M&V) of savings due to peak load reduction or energy efficiency measures (EEMs). The accuracy requirements of baselines and energy savings estimates depend on the particular application case. An owner wishing to track over-time performance relative to internal efficiency goals is likely to have much less stringent requirements than a regulated utility that offers incentive programs for reductions in building energy use.

Measurement and verification of energy savings can be conducted in a number of ways, as defined in the International Performance Measurement and Verification Protocol (IPMVP) [EVO 2012]. Savings may be determined based on isolation of a retrofit or efficiency measure, or based on more broadly encompassing measurements of whole-building meter use, before and after a measure. Although smaller savings are more easily obscured, and can be lost in the noise when using whole-building measurements, the advent of massively available interval meter data has enabled the development of more powerful baseline models, that those that have traditionally been used for whole-building performance characterization.

This baseline evaluation study was motivated by two trends in building energy efficiency. The first is the rapid growth in the number of commercially available energy information systems - an emerging technology that can enable up to 20% building energy savings, through continuous performance analysis, visualization, and feedback with an engaged user [Granderson 2009]. The second is a growing interest in whole-building focused utility programs, which in contrast to one-time single-measure interventions include, for example, EIS and continuous performance optimization, strategic energy management, and multiple-measure improvements. Although such programs have begun to be piloted, key barriers to scaled implementation are questions regarding the cost, accuracy, and time associated with the creation of whole-building baselines, and their use in energy savings calculations. Moreover, technologies such as EIS hold great promise, as they can enable deep energy savings, *and* automate the

creation of whole-building baselines; however the 'black-box' nature of proprietary baseline methods has raised even deeper questions concerning the robustness of the models and associated energy savings calculations.

The trends discussed above are leading to increasing use of baseline models for evaluating energy savings, so understanding the reliability of those models is important. This study established a preliminary methodology to evaluate the predictive accuracy of whole-building baseline models, using statistical performance metrics, and metered data from a test set of 29 buildings. This methodology was applied to evaluate the relative and absolute performance of five whole-building baseline models, ranging from simple to more sophisticated, and including a proprietary model from Pulse Energy, a commercial EIS provider. An important contribution 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.

2. Methodology

To assess the performance of each of the baseline models, a 4-step evaluation methodology was applied, as illustrated in Figure 1. This 4-step methodology represents a standard evaluation method called 'cross-validation', in which the model is fit using one set of data, the 'training data', and then used to predict data that were not included in fitting the model. Measures of model fit were then quantified, and compared. The Regional Technical Forum, California Public Utility Commission, and Portland Energy Conservation Inc. (PECI) contributed feedback on the methodology and study design, to encourage stakeholder acceptance.

In steps one and two, energy use predictions from several baseline models are generated, including multi-parameter change-point models, a mean-week model, the Pulse Adaptive Model, and the LBNL baseline regression model, each of which are described in Section 2.1. In steps three and four, the predictive ability of each baseline model is quantified, and the relative performance of each is evaluated.

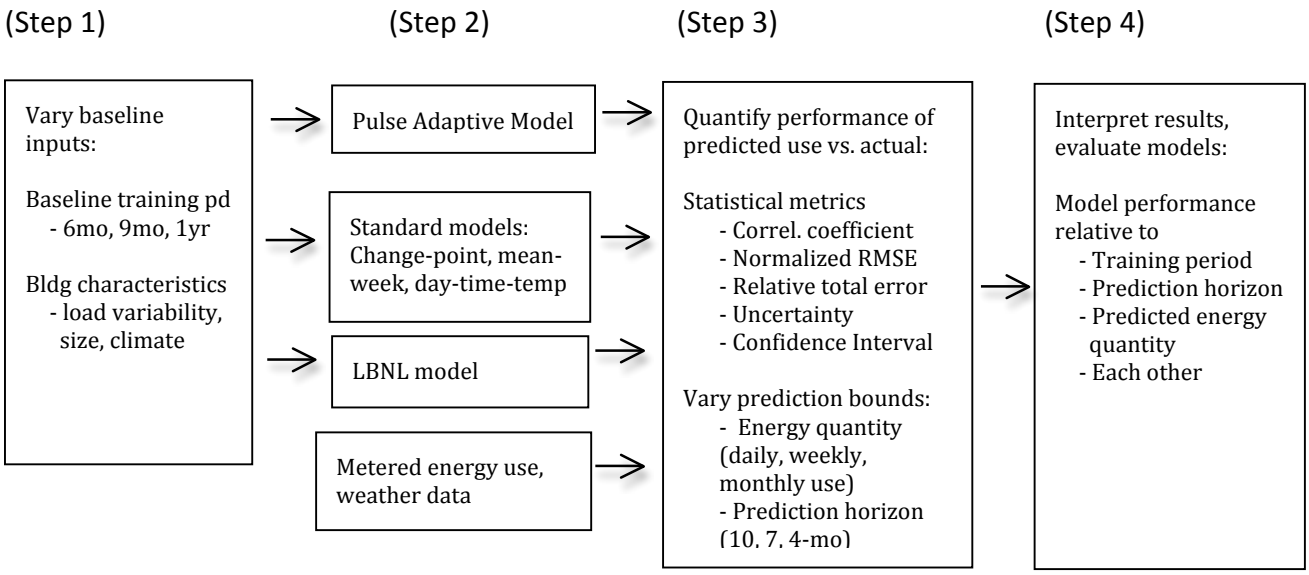

Figure 1: Schematic representation of the 4-step study methodology

In Step 1, two parameters were varied to determine the predictive performance of the baseline models across a diversity of conditions – the model training period, and the characteristics of the buildings included in the analysis.

1. Baseline training period - the amount of data used to build the model

Each model was trained using 6, 9, and 12 months of weather data and metered whole-building electric demand data. Given that the granularity of weather data is typically hourly, metered energy use was also considered at hourly intervals.

2. Building characteristics

16 months of metered energy data from 29 buildings was used in the analysis. For each set of data, according to all indications, no energy efficiency measures (EEMs) had been implemented, and the operation of the building was unchanged, representing a 'constant' set of energy use conditions. These buildings are located in a variety of climates and geographic locations. The set of 29 buildings is largely comprised of commercial offices, but does include a small number of non-office buildings. A summary of building characteristics is provided in Appendix B, for sites where the information was available.

In Step 2, baseline models and predictions were generated. Weather data was combined with meter data from the training period, to fit each baseline model. Once the fit of the model was determined, weather data from the prediction period was then used to generate model-predicted energy use for each building. For the Pulse Adaptive Model, the baselines and associated predictions for each building were generated using an application provided by Pulse Energy. The application inputs included weather data and meter data from the training period, and weather data from the prediction period.

For the non-proprietary, industry-standard and LBNL models, LBNL used automated programming scripts used to create the baselines and predictions. The models evaluated in the study are detailed in Section 2.1.

In Step 3 the performance of each baseline model was characterized according to a number of statistical performance metrics. The prediction bounds were varied according to two parameters:

1. The quantity being predicted

Models were developed from hourly interval meter and weather data, and then aggregated into daily, weekly, and monthly energy use predictions.

2. The prediction horizon - how far into the future predictions are made

16 months of metered energy use data was available for each of the 29 buildings included in the evaluation. Since the training periods were fixed to 6, 9, and 12 months, the associated prediction horizons were 10, 7, and 4 months, respectively.

The specific performance metrics that were used in the evaluation are detailed in Section 2.2.

In Step 4 of the evaluation methodology, the performance of each model was interpreted according to the set of statistical metrics computed in Step 3. The individual performance of each baseline method was evaluated, and the results were used to compare the methods relative to one another. The study parameters that were explicitly considered in the performance evaluation were the baseline training period, the prediction horizon, and the unit of prediction, i.e., daily, weekly, and monthly energy use. Due to limitations on the number of cases that could be analyzed under the scope of the study, hourly predictions were not considered.

2.1 Baseline Models

Five baseline models were evaluated in this study. These five models represent a sample of proprietary commercial methods, and public domain industry-standard and laboratory-developed methods that are used in the industry today. The mean-week model, which assumes that every week is the same in terms of energy use as a function of time, is the simplest model that might be used in practice. The change-point model and the day-time-temperature regression model are commonly used for M&V. The LBNL model is a more sophisticated regression model than those normally used for M&V and is thought by LBNL researchers to perform better than most models currently in use. The mean-week, change-point, day-time-temperature, and LBNL models all conform to requirement 5.2.10 of ASHRAE Guideline 14 [ASHRAE 2002], discussed in Section 4.2

Change-point (CP) models were the industry-standard before the advent of widely available interval meter data, and therefore do not include time. Detailed in [ASHRAE 2002; Haberl 2005], these models relate energy use to ambient temperature, according to a piecewise-continuous temperature response, with up to three temperature ranges.

For this study, the change points were chosen by optimization, allowing up to five temperature ranges, each with its own temperature response. The change point temperatures were determined so as to minimize the predictive error (not necessarily to represent physical significance), subject to the constraint that change points must be at least 4 F apart.

In the every-week-the-same model, also known as the *mean-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.

The *day-time-temperature (DTT) regression model* includes time of day, day of week, and two temperature variables to allow different heating and cooling slopes. The temperature variables are defined as the number of degrees F below 50 F, and the number of degrees above 65 F. The use of time-of-day and day-of-week variables is described in [Energy and Environmental Economics 2011], in the context of more complicated regression models that include special handling of, e.g. humidity and holidays.

The *Pulse Adaptive Model (PAM)* is a proprietary algorithm that is included in Pulse Energy's commercially available energy information system (EIS), which uses weighted averaging [Granderson 2009].

As described in [Mathieu 2011], the *LBNL model* is a regression model that includes time of week, and a piecewise-continuous temperature response with fixed change points that were set to 45, 55, 65, 75, and 85 F. Separate regressions were fit for 'occupied' and 'unoccupied' periods of the day. In [Mathieu 2011] the 'occupied' and 'unoccupied' periods were determined by visual inspection and used for determining different temperature behavior for the two modes. For the current analysis the determination of unoccupied and occupied periods was made by fitting a simple model similar to the day-time-temperature model: a day of the week was defined to be 'occupied' if most of the residuals from the simple model were positive (i.e., the building used more energy than predicted), otherwise it was defined as 'unoccupied'.

2.2 Baseline Model Performance Metrics

The statistical performance metrics that were considered in the study are collectively referred to as 'goodness-of-fit' metrics. They are each described in the following:

The *correlation coefficient* (r) quantifies the extent to which high predictions are associated with high data values, and low predictions are associated with low data values. A value of one indicates that predictions and data are perfectly related by a linear transformation, whereas a value of zero indicates no linear relationship between the predictions and the data. A value of negative one indicates that the data and the predictions are perfectly related by a linear transformation, but that high predicted

values map to low data values, and vice versa. A r value does not necessarily indicate accuracy: if the predictions are exactly equal to 10 times the data, plus 1000, they are very inaccurate but the correlation is perfect. The equation for r is provided in Equation 1, where E_i is the actual metered energy use per unit time, $\hat{E}_i$ is the model prediction, $\overline{E}$ is the mean of the metered energy use per unit time, $\overline{\hat{E}}$ is the mean of the model prediction, and n is the total number of predictions in the prediction horizon.

Equation 1: $$r = \frac{\sum_{i=1}^{n}\left(E_i - \overline{E}\right)\left(\hat{E}_i - \overline{\hat{E}}\right)}{\sqrt{\sum_{i=1}^{n}\left(E_i - \overline{E}\right)^2}\sqrt{\sum_{i=1}^{n}\left(\hat{E}_i - \overline{\hat{E}}\right)^2}}$$

The *root mean squared error* (RMSE) quantifies the typical size of the error in the predictions, in absolute units. In this study, mean kW was used as a convenient unit to avoid the fact that different months have different numbers of days; since power is energy per unit time, conversions between energy and power simply only required simple multiplication or division by a constant. The equation for RMSE is provided in Equation 2, where E_i is the actual metered energy use per unit time, $\hat{E}_i$ is the model prediction, and n is the total number of predictions in the prediction horizon[2].

Equation 2: $$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}\left(E_i - \hat{E}_i\right)^2}{n}}$$

The *normalized root mean squared error* (nRMSE) is the RMSE divided by the mean of the data. This metric also quantifies the typical size of the error, but does so relative to the mean of the data; for instance, a value of 0.1 means errors are typically about 10% of the mean value. Note that this is the same metric that ASHRAE 14 refers to as 'CV(RMSE)' [ASHRAE 2002]. The traditional statistical definition of 'coefficient of variation', or CV, is the standard deviation of a set of numbers, divided by the mean of that set of numbers. However, in the ASHRAE definition, the denominator is the mean of the *energy data*, rather than the mean of the *errors.* To avoid confusion with the traditional statistical terminology, this study uses the term 'normalized RMSE' rather than 'CV(RMSE)'. The equation for nRMSE is provided in Equation 3, where E_i is the actual metered energy use per unit time, $\hat{E}_i$ is the model prediction, and n is the total number of predictions in the prediction horizon.

[2] To quantify the RMSE of a model's predictions during the training period, i.e., relative to the data used to fit the model, the denominator of the equation is (n-p), where p is the number of parameters in the model. In contrast, the denominator is n when quantifying the model fit for the prediction period, as is the case with the cross-validation approach used in this study.

Equation 3: $$\text{nRMSE} = \frac{\sqrt{\frac{\sum_{i=1}^{n}\left(E_i - \hat{E}_i\right)^2}{n}}}{\frac{\sum_{i=1}^{n}\left(E_i\right)}{n}}$$

The *normalized mean absolute error* (nMAE) is the mean absolute error divided by the mean of the data. This metric is similar to nRMSE, but places less emphasis on extreme values. The equation for nMAE is provided in Equation 4, where E_i is the actual metered energy use per unit time, $\hat{E}_i$is the model prediction, and n is the total number of predictions in the prediction horizon.

Equation 4: $$\text{nMAE} = \frac{\frac{\sum_{i=1}^{n}\left|E_i - \hat{E}_i\right|}{n}}{\frac{\sum_{i=1}^{n}\left(E_i\right)}{n}}$$

The *relative bias* (relBias) is the mean of the predictions divided by the mean of the data. A value of 0.1 means that the prediction of the total energy used during the entire prediction horizon is 10% higher than the actual value; a value of -0.15 means the prediction is 15% lower. The equation for relBias is provided in Equation 5, where E_i is the actual metered energy use per unit time, $\hat{E}_i$ is the model prediction, and n is the total number of predictions in the prediction horizon.

Equation 5: $$\text{relBias} = \frac{\frac{\sum_{i=1}^{n}\left(\hat{E}_i\right)}{n}}{\frac{\sum_{i=1}^{n}\left(E_i\right)}{n}}$$

The *median relative total error (medRTE)* indicates whether the model has a *systematic* tendency to over- or under-predict. Suppose a model over-predicts one building by 10% (0.1), gets one exactly right, and under-predicts another by 10%. The relative total errors are thus 0.1, -0.1, and 0.0. The median of these is 0; that suggests that the modeling approach does not have an overall bias, but is not a good way of quantifying the typical error. The equation for medRTE is provided in Equation 6, where E_i is the actual metered energy use per unit time, $\hat{E}_i$ is the model prediction, and n is the total number of predictions in the prediction horizon.

Equation 6: $$\text{medRTE} = \text{median}\left\{\frac{\left(E_i - \hat{E}_i\right)}{E_i}\right\}$$

Relative to the median relative total error, the *median of the absolute relative total error (med(absRTE))* is a better metric to understand the typical error in the prediction of total energy use over the prediction horizon. The equation for med(absRTE) is given in Equation 7, where E_i is the actual metered energy use per unit time, $\hat{E}_i$ is the model prediction, and n is the total number of predictions in the prediction horizon. Continuing the example above, suppose a model over-predicts one building by 10% (0.1), gets one exactly right, and under-predicts another by 10%. The *absolute* relative total errors are thus 0.1, 0.0, and 0.1, and the median is .1, or 10%.

Equation 7: $$\text{med(absRTE)} = \text{median}\left\{\frac{\left|E_i - \hat{E}_i\right|}{E_i}\right\}$$

The final metric, *quantiles of residuals* (2.5%,10%, 50%, 80%, 97.5%), are helpful in characterizing the statistical distribution of the residuals, rather than just their typical size, as is true of the mean and median error metrics.

2.3 Study Validity and Confidence Intervals

Pulse Energy provided some of the 16-month metered building data that was used in study. To maintain study rigor and validity, it was necessary to verify that Pulse Energy did not artificially alter the predictions that they generated, in the cases for which they provided the study data, as they could have adjusted their predictions to provide a better match. (For the data sets that LBNL contributed to the study, Pulse Energy did not have access to meter data from the prediction horizon.) Therefore, the study methodology included an analysis to confirm that the relative total error and normalized root mean squared error were no better for the buildings provided by Pulse Energy than for the others.

Pulse Energy also provided the research team with 90% confidence intervals for the predictions of the Pulse Adaptive Model. Therefore, the research team also validated whether in fact the metered data was truly within the (prediction +/- confidence interval), 90% of the time.

3. Results

The presentation of results from the comparative model assessment focuses on nRMSE and median absolute relative total error. These are the metrics most critical to understanding the uncertainty in measurement and verification of energy savings, using baseline models. However, summary tables for the full set of performance metrics considered in the study are provided in Appendix A, including quantiles of residuals.

3.1 Normalized Root Mean Squared Error

The normalized root mean square error for each model, predicted energy quantity, and training period are summarized in Table 1. Since this metric quantifies the typical size of the error relative to the mean of the data, a value of 0.1, for example, indicates that errors are typically about 10% of the mean value.

Table 1. nRMSE for each model, predicted quantity, and training period

Predicted Quantity	Relative Performance	6-mo training period, 10-mo energy prediction	9-mo training period, 7-mo energy prediction	12-mo training period, 4-mo energy prediction
Daily Energy Use	Best	PAM (.16)	DTT (.17)	LBNL (.13)
		DTT (.18)	LBNL (.17)	PAM (.14)
		LBNL (.19)	MW (.18)	DTT (.17)
		MW (.20)	PAM (.19)	MW (.18)
	Worst	CP (.25)	CP (.22)	CP (.24)
Weekly Energy Use	Best	DTT (.13)	DTT (.10)	LBNL (.10)
		LBNL (.13)	PAM (.13)	DTT (.12)
		PAM (.13)	LBNL (.14)	PAM (.12)
		MW (.16)	CP (.15)	MW (.13)
	Worst	CP (.17)	MW (.15)	CP (.17)
Monthly Energy Use	Best	LBNL (.09)	DTT (.08)	LBNL (.08)
		DTT (.10)	LBNL (.09)	PAM (.10)
		PAM (.10)	PAM (.10)	DTT (.11)
		MW (.14)	CP (.11)	MW(.12)
	Worst	CP (.14)	MW (.13)	CP (.18)

Overall, the LBNL, Pulse Adaptive Model, and day-time-temperature model have smaller errors than that the mean-week and change-point models. The differences between the three best models are quite small, on the order of a percentage points. Across the entire data set, the nRMSE from the Pulse Adaptive Model range from 10-19% of the mean; nRMSE from the LBNL model range from 8-19% of the mean, and the nRMSE from the DTT model range from 8-18% of the mean. Monthly energy use was predicted with the least error, and daily energy was predicted with the most error. As the length of the prediction period decreases, the error in predicted energy use decreases; however, again, the differences were small, only on the order of a couple of percentage points for any given model.

Another useful view into the data, is to consider pair-wise comparisons of nRMSE for the different models in the study, with each of the 29 buildings explicitly represented as a single point. Points that fall directly on the 45-degree line indicate cases in which the error is the same for both models; points above or below the line indicate cases in which one model had higher or lower error than the other. Points near the lower left corner indicate buildings for which both models resulted in smaller predictive errors, while those near the upper right correspond to higher predictive errors. Any two models can

be compared using such plots; in Figures 2 through 4, the nRMSE for the LBNL, DTT, and change-point models are plotted relative to the Pulse Adaptive Model.

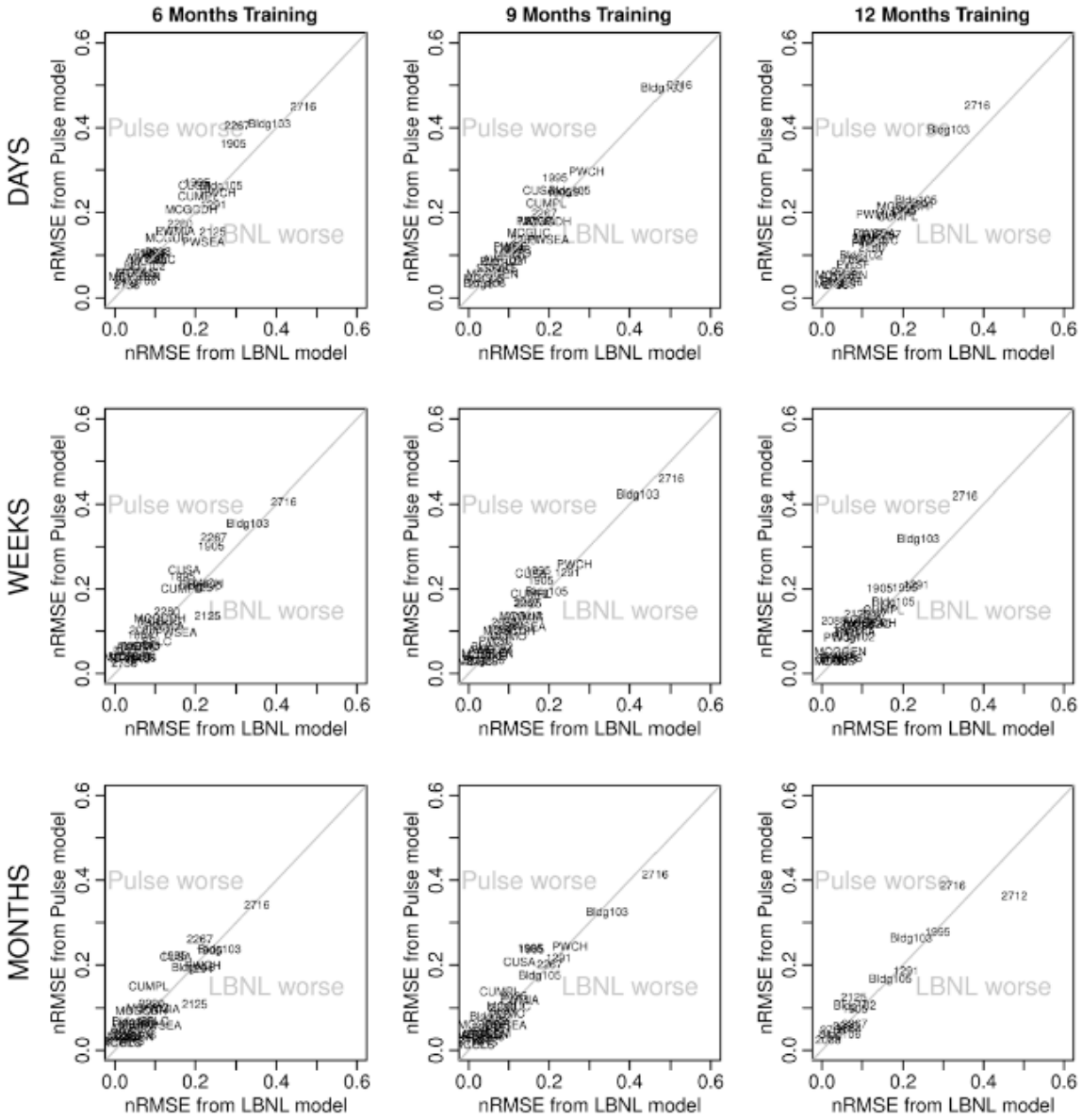

Figure 2: nRMSE of the energy predictions from the Pulse Adaptive Model vs. the LBNL model: rows show daily, weekly, monthly energy, and columns show 6-month, 9-month, and 12-month training periods.

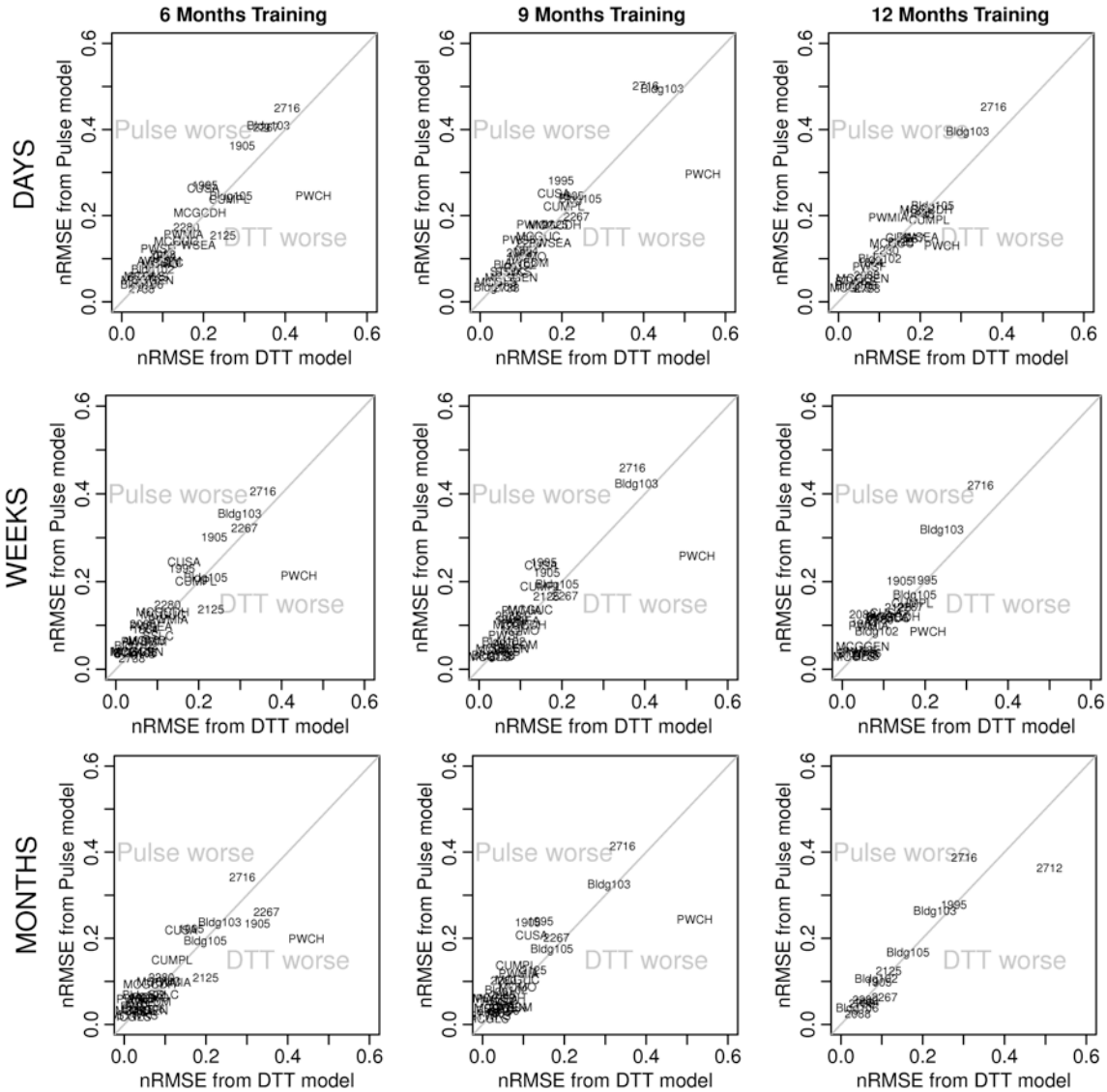

Figure 3: nRMSE of the Pulse Adaptive Model vs. the day-time-temperature (DTT) model – daily, weekly, and monthly energy use predictions for 6-month, 9-month, and 12-month training periods.

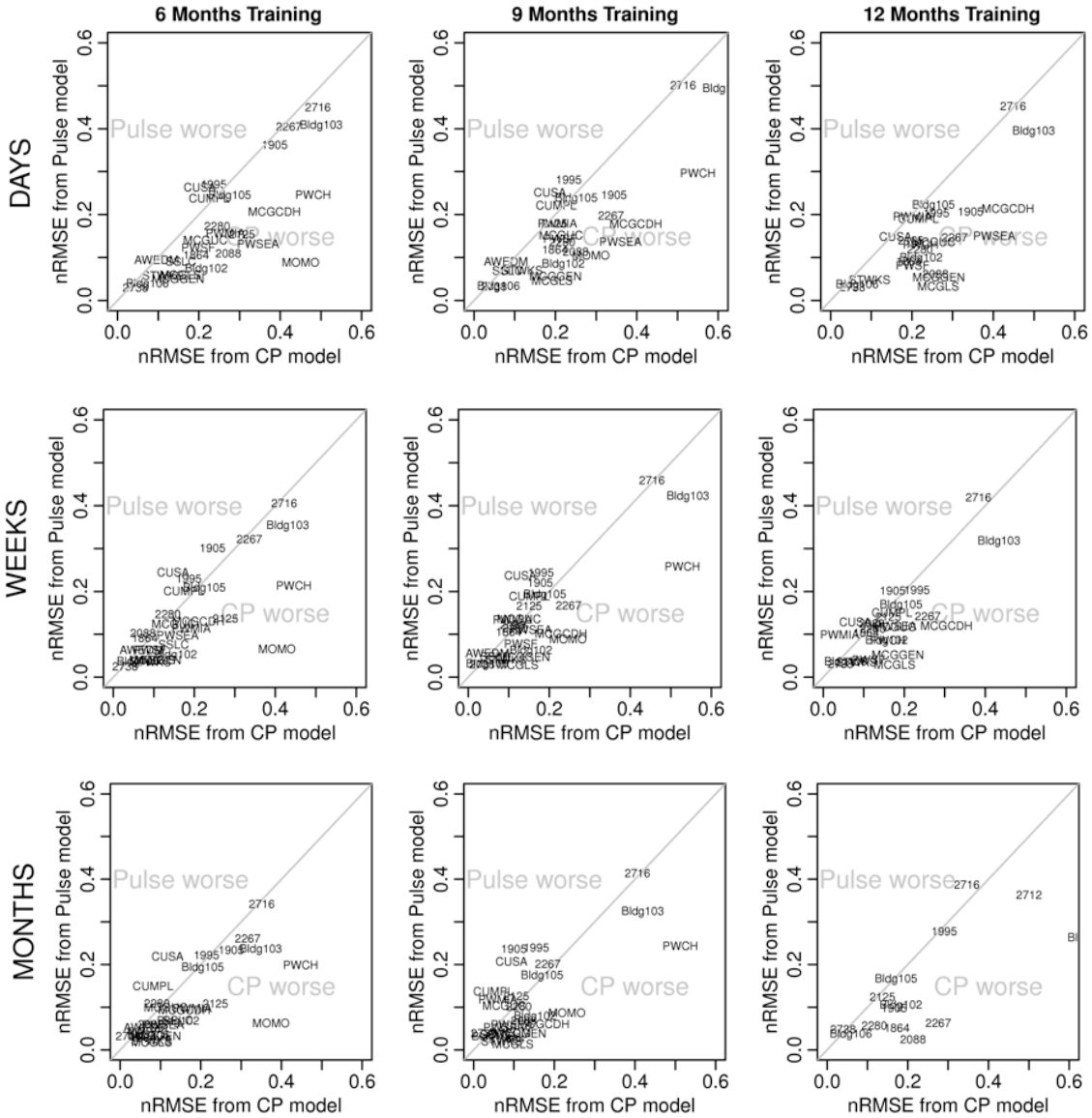

Figure 4: nRMSE of the Pulse Adaptive Model vs. the change-point (CP) model – daily, weekly, and monthly energy use predictions for 6-month, 9-month, and 12-month training periods.

3.2 Median Absolute Relative Total Error

The median absolute relative total error for each model, training period, and prediction horizon is summarized in Table 2. To compute this metric, the percent difference between the total predicted energy use (for the entire prediction period), and the actual energy use is determined, and the absolute value taken. This is done for each building in the study, and the median value reported. Therefore, a value of .04 for example, would indicate that the median error in total predicted energy use across the set of 29 buildings, was 4% of the actual energy use.

Table 2. Median Absolute Relative Total Error for Each Model, Training Period, and Unit of Prediction

Relative Performance	6-mo training period, 10-mo energy prediction	9-mo training period, 7-mo energy prediction	12-mo training period, 4-mo energy prediction
Best	MW (.030)	DTT (.039)	PAM (.035)
	LBNL (.032)	MW (.049)	LBNL (.041)
	DTT (.034)	CP (.052)	DTT (.046)
	PAM (.037)	LBNL (.055)	CP (.065)
Worst	CP (.051)	PAM (.061)	MW(.065)

Across all of the models, training periods and prediction periods, the median absolute percent error in predicted energy use ranged from 3% to 7%. Interestingly for M&V applications, the total error was smallest for a shorter training period and a longer prediction horizon, and was largest when the training period was much longer than the prediction horizon.

The relative performance of each model was mixed, varying with the training period and prediction horizon. Across the three different training and prediction periods, median absolute percent errors for the models were:

Day-time-temperature model, 3-5%
LBNL model, 3-6%
Pulse Adaptive Model, 4-6%
Mean-week model, 3-7%
Change-point model, 5-7%

3.3 Correlation

The error metrics discussed in 3.1 and 3.2 were concerned with how accurately daily, weekly, and monthly, and total energy use could be predicted. The performance metrics also included the correlation between predictions and data (see Appendix A). Correlation is a measure of the extent to which high predictions correspond to high data values and low predictions correspond to low data values. Correlation is not the same as accuracy, and therefore a less critical metric for this study. For example, if the predictions were always twice as high as the data, the accuracy would be poor, but the correlation would be perfect.

The results show that the Pulse Adaptive Model, LBNL, and DTT models all vastly outperform the change-point model and the mean-week model by this measure. Metered energy use was indeed high for the months, weeks, and especially the days for which these models predicted high consumption.

3.4 Study Validity and Pulse Adaptive Model Confidence Intervals

3.4.1 Study Validity

Figure 5 plots the Relative Total Error from the Pulse Adaptive Model versus the Relative Total Error for the LBNL model, with colors indicating which building data were provided by Pulse Energy and which were from other sources. Had Pulse Energy artificially

improved the model predictions in the cases where they provided the data, and thus had access to the meter data from the prediction period, the error from the Pulse Adaptive Model would be closer to 0 for the buildings that Pulse Energy provided, than for the others. That is, the model predictions would be systematically better than for the data that Pulse Energy did not provide. That is not in fact the case: when the LBNL model does poorly on this prediction, so does the Pulse Adaptive Model (points in the plot fall along the diagonal line), and to the same extent whether Pulse Energy knows the right answer or not (data from Pulse Energy and the other sources are equally distributed above and below the diagonal line). Similar investigations were conducted using the nRMSE metrics, and there is no evidence that the validity of the study was compromised.

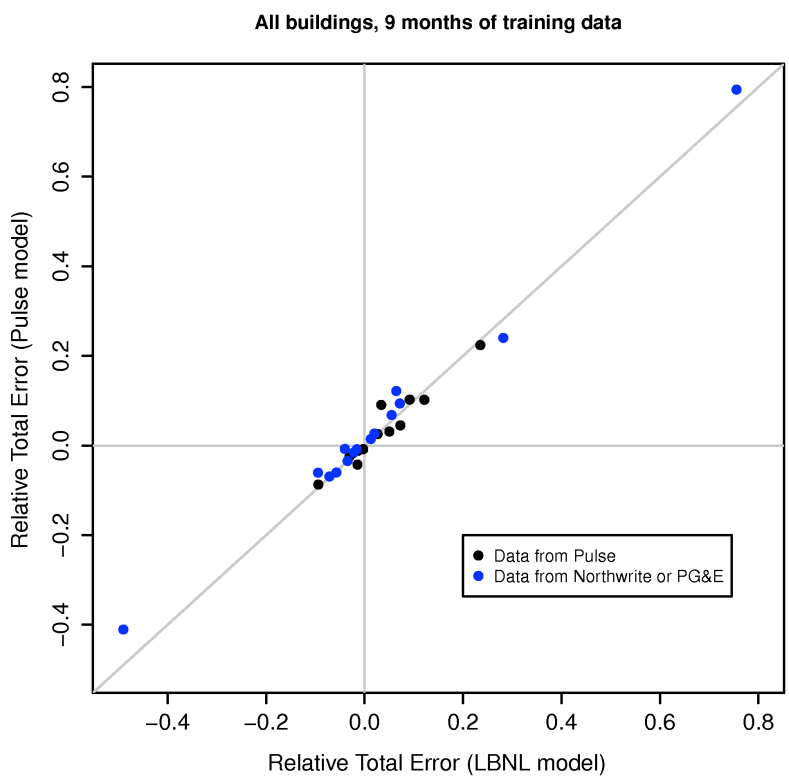

Figure 5: Relative total error for the LBNL versus Pulse models for 9 months of training data.

3.4.2 Confidence Interval

Although confidence intervals for any of the models could have been generated, the study scope prevented doing so. However, Pulse Energy does provide 90% confidence intervals for predictions from the Pulse Adaptive Model.

For each building, each amount of training data, and each unit of energy prediction, the fraction of predictions over the upper bound or below the lower bound was calculated. The histograms in Figure 6 show the fraction of days, weeks, or months under the lower bound or above the upper bound, for the Pulse Adaptive Model, for the case of 9

months of training data. For the 90% confidence interval, each histogram would ideally show that for every building, 5% of daily, weekly, or monthly energy use was below the lower bound and another 5% was above the upper bound. Systematically, however, more than 5% of actual data values exceed the upper bound, and more than 5% are less than the lower bound. This is true whether the predictions are for days, weeks, or months.

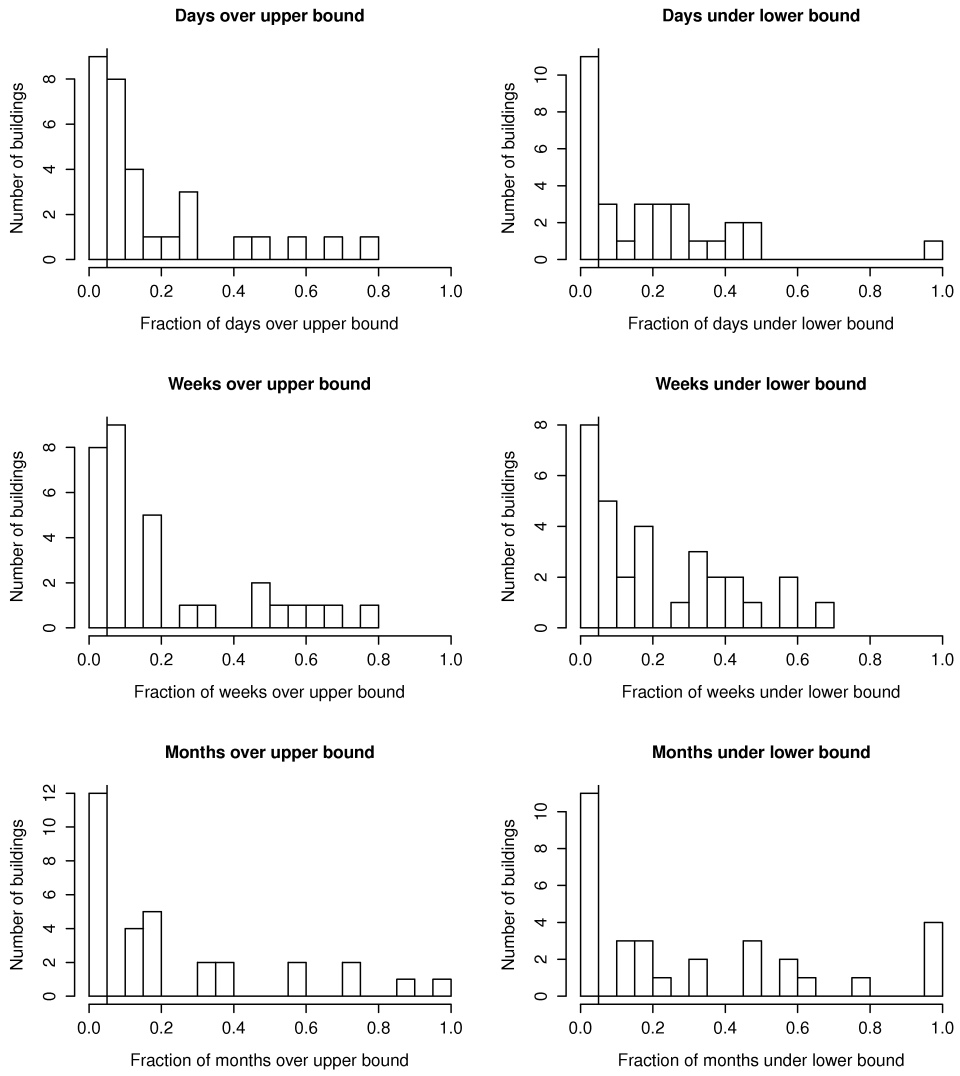

Figure 6: Histograms to assess the accuracy of reported confidence intervals for predictions from the Pulse Adaptive Model, using 9 months of training data.

Across each of the training periods analyzed, rather than 5%, 10% of the daily energy predictions exceeded the upper bound of the confidence interval, and 11% were below the lower bound. For the weekly predictions, 11% exceeded the upper bound, and 12% were below the lower bound. The difference was greatest for monthly energy, where

12% of the predictions exceeded the upper bound, and 15% were below the lower bound.

It is rather difficult to precisely identify confidence intervals, so a mild systematic error might be expected. The results do show that Pulse Energy's confidence intervals are too narrow: systematically, more than 5% of actual data values exceed their high bound, and more than 5% are lower than their lower bound. This is true whether the predictions are for days, weeks, or months. Interestingly, if the claimed confidence bounds were 80% bounds rather than 90% bounds, these medians would be very close to correct. Of course, for any individual building the results can differ greatly from the median. In fact, for some extreme cases as many as 75% of the predictions are below the lower bound, and for others as many as 75% of the predictions exceed the upper bound.

Although the confidence bounds are much too narrow for many individual buildings, and much too wide for some others, on average the approach correctly identifies which buildings need wide confidence bounds, and which buildings need narrow ones.

4. Discussion

In interpreting the study results, the focus of the discussion is the relative performance of the five baseline models according to key fitness metrics, and the extent to which the Pulse Adaptive Model is compliant with ASHRAE Guideline 14.

4.1 Relative Model Performance

Given the limited dataset and the non-representative nature of the data, widely generalizable conclusions cannot be drawn. However, the results of the study do indicate that the Pulse Adaptive Model, LBNL model, and day-time-temperature (DTT) models perform very similarly relative to the statistical metrics considered in this evaluation. They tend to out-perform the change-point and mean-week models, but on average performed equally well relative to one another. Were the change-point models used in this study limited to *fewer* than five temperature ranges, as in the standard modeling case (see Section 2.1), the change-point models would likely have performed slightly worse than in this study, leading to an even larger performance advantage for the Pulse and LBNL models.

There are several sources of temporal variation in building load that provide insight into the performance, or fit of each model.

1. Daily or weekly periodicity; this is very large for most buildings.
2. Temperature-dependence; this is small but not negligible for most buildings.
3. Other variation not explained above; this is small for most buildings, but moderate for others and very large for a few.

The mean-week model captures *only* number 1, the regular variation from one hour to the next within a week. The change-point model captures *only* number 2, the temperature-dependence. Over time, the daily and weekly periodicity cancel out, so

while the change-point model performs decently for long aggregation periods such as months, it performs very poorly for individual days. In contrast, the LBNL model, day-time-temperature model, and Pulse Adaptive Model capture number 1 and number 2. None of the models capture number 3, or can hope to do so, since this is variation that is not predicted by any explanatory variable available to the model. This is why the LBNL model and the and Pulse Adaptive Model both perform poorly on those buildings whose energy use varies in ways that aren't predictable from the outdoor temperature or the time of the week. The current dataset is too small to determine whether certain building sizes or types tend to have more unpredictable energy usage.

Although the change-point and mean-week models performed substantially worse than the others on average, even the mean-week model performed surprisingly well in an *absolute* sense. For instance, when a 12-month training period was used, and predictions were for monthly energy consumption the median nRMSE for the mean-week model was 12%, but the Pulse Adaptive Model and LBNL model had errors nearly as large, at 10% and 8%. The mean-week model also fared within a few percentage points of the other models in terms of median percent error in total predicted energy use.

It does not appear, however, that the somewhat poorer performance of the mean-week model is simply a statistical artifact due to the small sample size - unlike the difference between the LBNL and Pulse Adaptive Model, which might be. One indication of the inferiority of the mean-week model is that the poorer performance carries across all lengths of training periods and over predictions for days, weeks, and months. Conversely, the LBNL, DTT, and Pulse Adaptive models often switch ranks on the various metrics and various analyses. Furthermore, the mean-week model does not contain any temperature information at all, so there is a strong expectation that it will not perform as well as the other models. The change-point model also does not perform as well as the Pulse Adaptive Model, LBNL, and DTT models.

In terms of the median absolute percent error in total energy use over the full prediction horizon, the relative performance of each model was mixed, and depended on the length of the training period and prediction horizon. The difference between the LBNL, Pulse Adaptive Model and DTT models was only a couple of percentage points, and typical errors ranged from only 3-6% across the three cases considered.

In addition to the models' ability to accurately predict daily, weekly, monthly, and total energy use, the study also evaluated correlation between model predictions and metered data. Again, LBNL, Pulse Adaptive Model, and DTT models all vastly outperformed the change-point and mean-week models - the months, weeks, and especially days that for which the models predicted high energy use did indeed have high metered energy use, and vice versa.

4.2 Compliance with ASHRAE Guideline 14

ASHRAE Guideline 14 [ASHRAE 2002] defines two quantitative requirements for whole-building M&V:

1. Guideline 5.2.10 requires a 'net determination bias' less than 0.005%. Net determination bias is defined as the sum of the prediction errors divided by the sum of the load data (and multiplied by 100 to make a percent), where the sum is over the entire baseline period.
2. Guideline 5.3.2.1e states "The baseline model shall have a maximum CV(RMSE) of 20% for energy use and 30% for demand quantities when less than 12 months of post-retrofit data are available for computing savings. These requirements are 25% and 35%, respectively, when 12 to 60 months of data will be used in computing savings." .

The first requirement is extremely restrictive. It severely limits the range of approaches that can be used for creating baselines. The LBNL, mean-week, change-point, and day-time-temperature models all meet this criterion, which was one factor in including them in the study. LBNL did not have access to predictions from Pulse Energy's training period, and as such was not able to validate compliance. One limitation of this requirement is that even reasonable modifications that would likely improve model performance could result in non-compliance. For instance, since building performance changes with time, it might make sense to give more statistical weight to later data points than to earlier ones, however Guideline 5.2.10 essentially requires all points to be weighted equally. This requirement is particularly restrictive, considering that an alternative method for creating baseline predictions – calibrated whole-building simulation – allows 5% bias, a factor of 1000 higher than is allowed with statistical approaches.

Both requirements one and two are based on comparing the model's predictions to the data that *were* used to fit the model. As such, neither is amenable to independent validation in a strict sense: a modeler can always adjust their predictions post facto to make them comply with the guidelines. Although LBNL did not have access to predictions from Pulse Energy's training period, and therefore was unable to validate compliance, we note that when applied to data that were *not* used to fit the model, the Pulse Adaptive Model is no more biased than the other models.

Although it was not possible to independently test whether the Pulse Adaptive Model complies with these requirements, the study results indicate that in general, the Pulse Adaptive Model predicts energy use more accurately than change-point models. In turn, well-fit change-point models are one of the best-practice modeling approaches referenced in Guideline 14. The ability to accurately predict energy use beyond the training period is in many respects, a more relevant test of the real-world usefulness (and lack of bias) in the model. While the Guideline 14 requirement 5.3.2.1e refers to the training period, *for the prediction period*, the LBNL model, day-temperature-time

model, and Pulse Adaptive Model all comfortably meet required threshold for most of the buildings in this data set; all fail to meet the requirement for the most unpredictable buildings.

5. Conclusion

This study leveraged the data from 29 buildings, spanning a diversity of climates, sizes, and geographical locations, to evaluate the uncertainty in energy use predictions from five baseline models. The methodology developed and applied for this study comprises the first step in determining the uncertainty in the measurement and verification (M&V) of gross, whole-building energy savings.

Relative to the statistical performance metrics that were evaluated in this study, the Pulse Adaptive Model, LBNL model, and day-time-temperature models all outperformed the change-point models. This has important implications for whole-building M&V, as change-point models are the traditional industry-standard whole-building modeling approach. However, since the dataset for this study was limited to a small number of buildings, and not fully representative, widely generalizable conclusions cannot be established.

6. Future Work

Future work will focus on expansion of the number of datasets, and diversity of buildings to enable more generalizable findings, disaggregation of different building types, and correlation of errors with load variability. In addition, including more proprietary models will provide further insights into the robustness of commercial tools that promise to automate the creation of baselines for purposes such as continuous anomaly detection, and whole-building measurement and verification.

The uncertainty in the measurement and verification of whole-building energy savings that are calculated in a manner commensurate with IPMVP Option C [EVO 2012] depends on two key factors:

1. The accuracy of model predictions of building energy use, according to the operation of the building during the model training period, *prior* to the energy efficiency measures. This is the accuracy that was addressed in this study.
2. The degree to which, excluding all effects but those associated with the EEMs, the building's operation and energy use during the baseline period, is equal to that of the post-measure period.

This study considered only the first factor; future work will address the second factor as well, requiring data from a number of buildings in which EEMs have been implemented. This will entail permit quantification of the uncertainty in both the baseline model, as well as the energy savings that the baseline is used to calculate.

Acknowledgement
This work was supported by the Assistant Secretary for Energy Efficiency and Renewable Energy, Building Technologies Program, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

The authors would like to acknowledge Pulse Energy for supporting this study, and David Helliwell, Harish Raisinghani and Bruce Herzer, in particular. In addition, the authors thank LBNL's Demand Response Research Center, and Bill Koran of NorthWrite, for contributing a portion of the building data used this study. Without a sufficient volume and diversity of data, meaningful insights would not have been possible.

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.

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

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.

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.

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

Appendices

Appendix A: Detailed Statistical Performance Metrics

Detailed statistical performance metrics for each of the baseline models are provided in Tables A1- A10. Tables A1-A3 summarize the median model performance for daily energy use predictions, Tables A4-A6 correspond to weekly energy predictions, and Tables A7-A9 correspond to monthly energy predictions. Table A10 summarizes the relative total error for each model, training period, and prediction horizon.

Table A1. Median model performance over entire data set, daily energy predictions, 6-month training period, 10-month prediction horizon

Model	r	nRMSE	relMAE	relTotErr	2.50%	10%	50%	90%	97.50%
PAM	0.765	0.157	0.116	0.011	-0.198	-0.125	0.013	0.151	0.398
LBNL	0.779	0.189	0.131	-0.008	-0.2	-0.115	-0.007	0.169	0.326
DTT	0.716	0.178	0.124	0.012	-0.203	-0.113	0.009	0.178	0.312
CP	0.189	0.245	0.206	0.016	-0.382	-0.235	0.003	0.321	0.446
MW6	0.483	0.204	0.162	0.009	-0.324	-0.174	0.008	0.174	0.297

Table A2. Median model performance over entire data set, daily energy predictions, 9-month training period, 7-month prediction horizon

Model	r	nRMSE	relMAE	relTotErr	2.50%	10%	50%	90%	97.50%
PAM	0.743	0.178	0.12	0.026	-0.196	-0.125	0.012	0.189	0.325
LBNL	0.753	0.172	0.133	0.013	-0.204	-0.114	0.011	0.183	0.334
DTT	0.746	0.171	0.129	0	-0.211	-0.118	-0.004	0.183	0.302
CP	0.006	0.217	0.186	-0.013	-0.328	-0.223	-0.042	0.285	0.389
MW	0.558	0.177	0.124	-0.012	-0.222	-0.113	-0.008	0.132	0.207

Table A3. Median model performance over entire data set, daily energy predictions, 12-month training period, 4-month prediction horizon

Model	r	nRMSE	relMAE	relTotErr	2.50%	10%	50%	90%	97.50%
PAM	0.731	0.144	0.108	-0.008	-0.219	-0.146	-0.012	0.13	0.263
LBNL	0.795	0.13	0.099	-0.013	-0.196	-0.131	-0.007	0.14	0.247
DTT	0.741	0.167	0.119	-0.01	-0.176	-0.112	-0.006	0.158	0.279
CP	0.088	0.238	0.191	-0.036	-0.343	-0.294	-0.057	0.249	0.353
MW	0.743	0.18	0.153	-0.011	-0.195	-0.121	-0.02	0.097	0.146

Table A4. Median model performance over entire data set, weekly energy predictions, 6-month training period, 10-month prediction horizon

Model	r	nRMSE	relMAE	relTotErr	2.50%	10%	50%	90%	97.50%
PAM	0.489	0.131	0.103	0.011	-0.164	-0.092	0.01	0.139	0.229
LBNL	0.558	0.13	0.109	-0.008	-0.128	-0.097	0.004	0.129	0.196
DTT	0.524	0.125	0.099	0.012	-0.13	-0.085	0.008	0.109	0.202
CP	0.193	0.174	0.139	0.016	-0.228	-0.142	0.029	0.169	0.262
MW	0.289	0.164	0.13	0.009	-0.213	-0.118	0.011	0.156	0.243

Table A5. Median model performance over entire data set, weekly energy predictions, 9-month training period, 7-month prediction horizon

Model	r	nRMSE	relMAE	relTotErr	2.50%	10%	50%	90%	97.50%
PAM	0.349	0.134	0.104	0.026	-0.136	-0.104	0.015	0.128	0.208
LBNL	0.518	0.136	0.106	0.013	-0.152	-0.079	0.014	0.125	0.19
DTT	0.652	0.102	0.09	0	-0.14	-0.098	-0.004	0.114	0.169
CP	0.001	0.151	0.114	-0.013	-0.2	-0.135	-0.022	0.115	0.237
MW	0.293	0.152	0.108	-0.012	-0.132	-0.085	0.002	0.108	0.159

Table A6. Median model performance over entire data set, weekly energy predictions, 12-month training period, 4-month prediction horizon

Model	r	nRMSE	relMAE	relTotErr	2.50%	10%	50%	90%	97.50%
PAM	0.358	0.12	0.098	-0.008	-0.129	-0.091	-0.034	0.142	0.226
LBNL	0.491	0.101	0.08	-0.013	-0.104	-0.079	-0.012	0.077	0.148
DTT	0.64	0.117	0.087	-0.01	-0.074	-0.066	-0.011	0.093	0.128
CP	0.125	0.17	0.138	-0.036	-0.188	-0.14	-0.049	0.083	0.195
MW	0.396	0.132	0.108	-0.011	-0.085	-0.08	-0.014	0.054	0.116

Table A7. Median model performance over entire data set, monthly energy predictions, 6-month training period, 10-month prediction horizon

Model	r	nRMSE	relMAE	relTotErr	2.50%	10%	50%	90%	97.50%
PAM	0.699	0.1	0.087	0.011	-0.072	-0.055	0.017	0.09	0.124
LBNL	0.669	0.091	0.076	-0.008	-0.092	-0.065	0	0.067	0.126
DTT	0.687	0.096	0.078	0.012	-0.068	-0.052	0.01	0.104	0.116
CP	0.413	0.143	0.107	0.016	-0.142	-0.082	0.013	0.131	0.162
MW	0.139	0.141	0.118	0.009	-0.122	-0.102	0.005	0.115	0.162

Table A8. Median model performance over entire data set, monthly energy predictions, 9-month training period, 7-month prediction horizon

Model	r	nRMSE	relMAE	relTotErr	2.50%	10%	50%	90%	97.50%
PAM	0.478	0.104	0.096	0.026	-0.08	-0.057	0.012	0.072	0.129
LBNL	0.626	0.092	0.079	0.013	-0.061	-0.05	-0.004	0.06	0.084
DTT	0.603	0.082	0.076	0	-0.06	-0.046	-0.003	0.058	0.068
CP	-0.008	0.106	0.083	-0.013	-0.096	-0.078	-0.022	0.063	0.089
MW	0.495	0.127	0.102	-0.012	-0.076	-0.06	-0.007	0.054	0.067

Table A9. Median model performance over entire data set, monthly energy predictions, 12-month training period, 4-month prediction horizon

Model	r	nRMSE	relMAE	relTotErr	2.50%	10%	50%	90%	97.50%
PAM	0.892	0.104	0.096	-0.008	-0.095	-0.078	-0.029	0.074	0.094
LBNL	0.949	0.082	0.078	-0.013	-0.093	-0.071	-0.014	0.037	0.063
DTT	0.92	0.107	0.09	-0.01	-0.069	-0.045	-0.002	0.068	0.074
CP	0.401	0.183	0.143	-0.036	-0.108	-0.095	0.028	0.201	0.241
MW	0.908	0.124	0.118	-0.011	-0.08	-0.041	0.019	0.071	0.096

Table A10. Median Relative Total Error for each model, predicted quantity, and training period

Relative Performance	6-mo training period, 10-mo energy prediction	9-mo training period, 7-mo energy prediction	12-mo training period, 4-mo energy prediction
Best	LBNL (-.008)	DTT (.000)	PAM (-.008)
	MW (.009)	MW (-.012)	DTT (-.010)
	PAM (.011)	LBNL (.013)	MW (-.011)
	DTT (.012)	CP (-.013)	LBNL (.013)
Worst	CP (.016)	PAM (.026)	CP (.036)

The Median Relative Total Error metric is near 0 for all of the models, but all that shows is that none of the models have a *systematic* tendency to over- or under-predict future total energy use. Suppose a model over-predicts one building by 10% (0.1), gets one exactly right, and under-predicts another by 10%. The relative total errors are thus 0.1, 0.1, and 0.0. The median of these is 0; that suggests that the modeling approach does not have an overall bias, but is not a good way of quantifying the typical error.

Appendix B: Building Characteristics

Table A-11 summarizes the known characteristics for the buildings that were included in this study. Commercial building type is provided for 25 of the 29 buildings, and floor area for 14 of the 29 buildings.

Table A-11. Floor area and commercial type for buildings included in the study

Location	Area (sq.ft)	Building Type
Northern Alberta, CA	2,000	Restaurant
Southern British Columbia, CA	11,850	Mixed-use campus building
Southern British Columbia, CA	19,400	Mixed-use campus building
Southern Quebec, CA	97,450	Mixed-use campus building
Southern Quebec, CA	96,250	Mixed-use campus building
Southern Quebec, CA	200,000	Mixed-use campus building
Southern Quebec, CA	86,200	Mixed-use campus building
Southern British Columbia, CA	1,300	Restaurant
Southwestern North Carolina, US	15,622	Office building
Southern Florida, US	7,700	Office building
Western Washington, US	12,000	Office building
Northern CA, US	20,000	Office building
Southern British Columbia, CA	206,400	Sports Complex
Southern British Columbia, CA	15,670	Restaurant
Northwestern Oregon, US		Office building
Northwestern Oregon, US		Office building
Northwestern Oregon, US		K-12 school
Southeastern Minnesota, US		Sports Complex
Northwestern Oregon, US		Office building
Northern Central Colorado, US		University dormitory
Southern Idaho, US		K-12 school
Southern Idaho, US		K-12 school
District of Columbia, US		Office building
District of Columbia, US		Office building
Southern Idaho, US		Hospital