LBNL-XXXXX

ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

Development and Application of a Statistical Methodology to Evaluate the Predictive Accuracy of Building Energy Baseline Models

Jessica Granderson, Phillip N Price

Energy Technologies Area

July, 2013

This page intentionally left blank.

DISCLAIMER

This report was prepared as the result of work sponsored by the California Energy Commission. It does not necessarily represent the views of the Energy Commission, its employees or the State of California. The Energy Commission, the State of California, its employees, contractors and subcontractors make no warrant, express or implied, and assume no legal liability for the information in this report; nor does any party represent that the uses of this information will not infringe upon privately owned rights. This report has not been approved or disapproved by the California Energy Commission nor has the California Energy Commission passed upon the accuracy or adequacy of the information in this report.

This page intentionally left blank.

Development and Application of a Statistical Methodology to Evaluate the Predictive Accuracy of Building Energy Baseline Models

Jessica Granderson[†], Phillip N Price

Lawrence Berkeley National Laboratory
1 Cyclotron Rd., Berkeley CA, 94720, USA

[†] Corresponding Author: JGranderson@lbl.gov; 1 Cycltron Rd., MS 90-3111, Berkeley CA, 94720; (510) 486.6792.

Abstract
This paper documents the development and application of a general statistical methodology to assess the accuracy of baseline energy models, focusing on its application to Measurement and Verification (M&V) of whole-building energy savings. The methodology complements the principles addressed in resources such as ASHRAE Guideline 14 and the International Performance Measurement and Verification Protocol. It requires fitting a baseline model to data from a ``training period'' and using the model to predict total electricity consumption during a subsequent ``prediction period.''

We illustrate the methodology by evaluating five baseline models using data from 29 buildings. The training period and prediction period were varied, and model predictions of daily, weekly, and monthly energy consumption were compared to meter data to determine model accuracy. Several metrics were used to characterize the accuracy of the predictions, and in some cases the best-performing model as judged by one metric was not the best performer when judged by another metric.

Keywords: baseline model; prediction; measurement and verification; energy savings; performance accuracy; whole-building energy

1. Introduction

There is growing recognition that whole-building-focused approaches to energy efficiency hold great promise in realizing deep and persistent energy savings in commercial buildings. Owners, property and facility managers, and utility incentive programs are increasingly adopting and piloting multi-measure strategies that move beyond traditional component-based or one-time commissioning or retrofit interventions; the industry is seeing a movement toward continuous energy improvement practices that may include efforts such as ongoing commissioning, strategic energy management, and operational optimization, as well as retrofits and the implementation of advanced control and information technologies [1].

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) [2]. Savings may be determined based on isolation of a retrofit or efficiency measure, or based on more broadly encompassing measurements of metered whole-building energy use, before and after the improvement. In theory, whole-building approaches, which often combine energy management processes with efficient technologies, and information technologies, would naturally lend themselves to measured, whole-building savings quantification, as opposed to measure-specific or calculated savings. Historically, however, whole-building Measurement and Verification (M&V) has relied upon monthly utility data, and therefore monthly baseline models, in which smaller levels of savings could be easily obscured due to model error. The IPMVP recommends that whole-building savings quantification be applied in cases where the expected savings are greater than ten percent, and where at least twelve months of pre- and post-data is available [2].

Today, the advent of increasingly available interval meter data has enabled the development of more robust baseline models, than the monthly models that have traditionally been used to characterize whole-building energy performance. In addition, whole-building approaches to efficiency have the potential to generate deeper energy savings than single-measure approaches. Moreover, many of the technologies included in whole-building efficiency strategies, such as energy information systems (EIS) and ongoing commissioning systems, not only *enable* energy savings of up to twenty percent [3], but include baselining functionality that can be used to automatically quantify savings according to the principles of IPMVP Option C [4, 5].

Although whole-building efficiency programs, interval meter data, and enabling building information technologies hold great promise in realizing deep energy savings in the commercial buildings sector, several questions relating to savings quantification remain to be answered: What metrics should be used to quantify the performance of these tools? Does interval data reduce the time required for measurement and verification relative to that required when using utility billing data? How accurate are baseline models based on interval meter data? Are savings of approximately ten percent, still

required for an acceptable degree of certainty in reported savings? How can proprietary tools that automate gross M&V be evaluated? This paper documents research findings that begin to address these questions. We present a statistical methodology to evaluate the predictive accuracy of baseline energy models used for whole-building savings quantification, and apply the methodology to assess the performance of five specific models. These models range from simple to more sophisticated, and include a proprietary model included in a commercial EIS offering.

While resources such as the IPMVP and ASHRAE Guideline 14, establish procedural and quantitative requirements for baseline model construction, goodness of fit to data during the model training period, and rules of thumb for model application given different expected depths of savings, they do not provide a general means of assessing model performance during a *prediction* period. The 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 dozen 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 diverse time scales of interest.

This methodology shares important similarities to the approaches used in the ASHRAE 'shootouts' of the mid and late 1990s [6; 7]. 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 three buildings, and the cross-validation was conducted only during a short subset of the model training period. Also, the ASHRAE shootouts focused on hourly quantifications of energy use, whereas in this study we consider daily, weekly, and monthly energy predictions. The ASHRAE competitions considered total energy use from a sum of submetered quantities, but the demonstration in this study is limited to data and models of whole building electric metering: that is the only meter data that was available in our dataset, and is all that is readily available in most buildings .

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. In addition, it provides a general approach to evaluate the errors in calculated energy savings, according to diverse pre- and post-measure time horizons, and large test sets of building energy data.

2. Methodology

The methodology developed to assess baseline prediction accuracy comprises four steps, as illustrated in Figure 1. This 4-step methodology represents a statistical approach called 'cross-validation', in which the model is fit using one set of data, the

'training data', and then used to predict future consumption data that were not included in fitting the model. Measures of model fit are then quantified and compared.
In steps one and two, energy use predictions from baseline models of interest are generated. In steps three and four, the predictive ability of each baseline model is quantified, and the relative performance of each is evaluated.

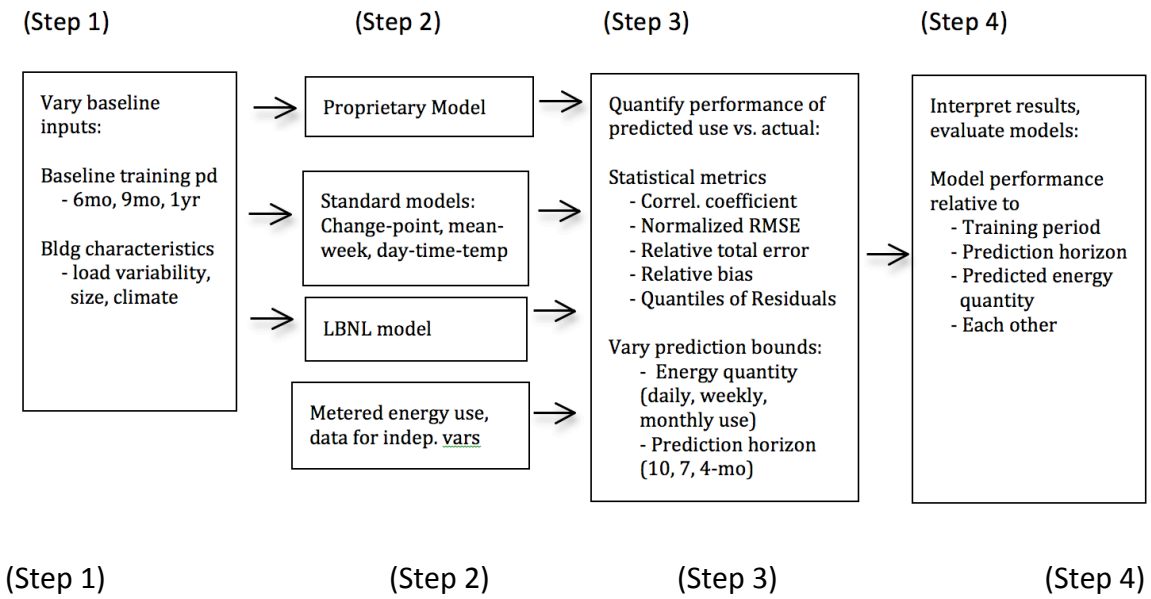

Figure 1: Schematic representation of the 4-step model evaluation methodology, and specific parameters used in demonstrating the methodology in this study.

In Step 1, two parameters are 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

In demonstrating the methodology, each model was trained using weather data and metered whole-building electric demand data from a 6-, 9-, and 12-month period. We used hourly outdoor air temperature data and hourly electricity consumption. In all cases the start of the training period was at the beginning of the building load data, which was January 1 in 14 of the buildings; various days in April in 12 of the buildings; and February, August, and September in the other three. In principle the entire analysis would be repeated using different start months in each building. For the dataset used in this report, which has only sixteen months of data available for most of the buildings, our ability to do this would have been very limited.

2. Building characteristics

16 months of metered electric data from 29 buildings was used in the demonstration of the methodology. 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 C, for sites where the information was available.

In Step 2, baseline models and predictions are generated. Metered energy consumption data and associated independent variables are used to fit each baseline model. Once the fit is determined, data for the independent variables, measured during the prediction period are used to generate model-predicted energy use for each building. The models used to demonstrate the methodology are detailed in Section 2.1; the independent variables they require include outside air temperature, and time of day or week.

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

1. The quantity being predicted

In demonstrating the methodology, 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

Prediction horizons are varied based on the length of data in the test data set, and the time periods of most interest for the evaluation application. In this demonstration, 16 months of metered energy use data were available for each of the 29 buildings. 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 is interpreted according to the set of statistical metrics computed in Step 3.

According to this construction, the model 'training' or 'fit' data is analogous to the pre-measure period in M&V applications, and the prediction period is analogous to the post-measure, or savings period. Model fitness measures to the metered data from the prediction period, then relate to the ultimate error incurred in applying the model to quantify savings, in the general case.

For this study, five baseline models were assessed according to the performance evaluation methodology, and the results were used to compare the models relative to one another, and in an absolute sense. The 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 chosen to demonstrate the methodology, and to explore questions of predictive accuracy under a set of varied training and prediction horizons. These five models include public domain methods commonly used in the industry, as well as a model developed by researchers at Lawrence Berkeley National Laboratory (LBNL), and a proprietary model used in a commercial EIS offering; a simplistic 'naïve' model was also included to serve as a comparative 'floor' on performance.

In 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 or each month in the year.

Change-point (CP) models were the industry-standard before the advent of widely available interval meter data, and do not include time. Detailed in [8; 9], 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 2 C (4 F) apart.

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 were defined as the number of degrees C below 10 C (50 F), and the number of degrees above 18 C (65 F). The use of time-of-day and day-of-week variables is described in [10], in the context of more complicated regression models that include special handling of, e.g., humidity and holidays.

The *proprietary model (Propr)* is offered in a commercially available energy information system (EIS). The model predicts the electric load at a given time as a weighted average of energy use at other times, giving higher statistical weight to data when times and conditions were similar to the given time than to other data, where "similarity" is defined according to a proprietary algorithm that takes into account the time of day, day of the week, outdoor air temperature, and other variables if provided [4].

The *LBNL model*, described in [11], is a regression model that includes time of week, and a piecewise-continuous temperature response with fixed change points that were set to 7, 13, 18, 24, and 29 C (45, 55, 65, 75, and 85 F). Separate regressions were fit for 'occupied' and 'unoccupied' periods of the day. The determination of unoccupied and occupied periods was made by fitting a linear regression model with two explanatory

variables, degrees below 10 C (50 F) and degrees above 18 C (65 F), and aggregating the results by time of the week: a time period 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 Differences between the models

There are several sources of temporal variation in building load that affect the performance of each model.

1. Daily and weekly periodicity (such as high load Wednesday afternoon, low load Sunday night); this is very large for most buildings, typically accounting for more than 70% of the variance in hourly load.
2. Temperature-dependence; this is small but not negligible for most buildings, accounting for 5-15% of the variance in hourly load.
3. Other variation not explained above, such as variation due to changes in occupant behavior, in the lighting or equipment used in the building, in the number of occupants, and so on. This variability is small for most buildings (less than 15%), but moderate for others and large (more than 50%) for a few.

The mean-week model captures *only* number 1, the regular variation of hourly load that occurs every week. The change-point model explicitly captures *only* number 2, the temperature-dependence; however, because the outdoor air temperature is higher during the day than at night, the hourly load predictions from this model also captures some of the daily load pattern in most buildings.

The LBNL model, day-time-temperature model, and proprietary model capture both 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.

2.3 Baseline Model Performance Metrics

The statistical performance metrics that are included in the baseline model evaluation methodology are collectively referred to as 'goodness-of-fit' metrics. Those most relevant to whole-building M&V applications, and referenced in ASHRAE Guideline 14 are described below; additional metrics also considered in development of the methodology are included in Appendix A.

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)' [8]. 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 1, 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 1: $$\text{nRMSE} = \frac{\sqrt{\frac{\sum_{i=1}^{n}\left(E_i - \hat{E}_i\right)^2}{n}}}{\frac{\sum_{i=1}^{n}\left(E_i\right)}{n}}$$

The *relative bias* (relBias) is the mean of the error in 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 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.

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

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. This metric is similar to the more commonly used "Mean Absolute Percent Error," but uses the median rather than the mean to quantify the central tendency. The median is less sensitive to extreme values which tend to arise from unusual or pathological cases. For instance, if the energy consumption during a single day is very low (which can happen due to a data recording error or an equipment malfunction), a small absolute error in the baseline prediction can lead to an enormous, or even infinite, relative error in the prediction. Calculating the mean relative total error would allow such days, if present, to substantially affect the error assessment, but the median is insensitive to them, so we prefer the median for this application. 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 3: $$\text{med(absRTE)} = \text{median}\left\{\frac{\left|E_i - \hat{E}_i\right|}{E_i}\right\}$$

The value of relBias is independent of the timescale being evaluated: if the average hourly energy use is too high by a factor of 1.1, then the average daily, weekly, monthly, and total energy use will also be too high by that same factor, so this is a measure of the total error. In contrast, the other metrics quantify the predictive accuracy at the timescale of the data and predictions: if the predictions and data apply to individual hours then these metrics summarize the error in the hourly predictions; for monthly predictions and data, they summarize the error in the monthly predictions. The numbers can be quite different. For example, imagine that every day half of the hourly predictions are too high by X kWh, and half are too low by X kWh. The root-mean-squared error in the hourly predictions would be X kWh, but the root-mean-squared error in the daily prediction would be 0 kWh.

3. Results

The results that are presented focus on comparative model assessment, using the metrics deemed most critical to understanding the error in measurement and verification of building energy savings, that is, nRMSE and median absolute relative total error. Summary tables for the full set of performance metrics considered in the methodology are provided in Appendix B.

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, which gives the median value when the model is fit to all of the buildings in the dataset. We use the median rather than the mean to avoid being influenced by extreme values: as shown in Figure 2, a few of the buildings are very poorly fit (by all of the models), and using the mean would let those few buildings dominate the error summary. 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. Median 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	PROPR. (.16)	DTT (.17)	LBNL (.13)
		DTT (.18)	LBNL (.17)	PROPR. (.14)
		LBNL (.19)	MW (.18)	DTT (.17)
		MW (.20)	PROPR. (.19)	MW (.18)
	Worst	CP (.25)	CP (.22)	CP (.24)
Weekly Energy Use	Best	DTT (.13)	DTT (.10)	LBNL (.10)
		LBNL (.13)	PROPR. (.13)	DTT (.12)
		PROPR. (.13)	LBNL (.14)	PROPR. (.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)	PROPR. (.10)
		PROPR. (.10)	PROPR. (.10)	DTT (.11)
		MW (.14)	CP (.11)	MW(.12)
	Worst	CP (.14)	MW (.13)	CP (.18)

Overall, the LBNL, proprietary, and day-time-temperature models had smaller errors than that the mean-week and change-point models. The differences between the three best models were quite small, on the order of a percentage points. Across the entire study set, the nRMSE from the DTT model ranged from 8-18% of the mean, the nRMSE from the LBNL model ranged from 8-19% of the mean, and the nRMSE from the proprietary model ranged from 10-19% of the mean. Monthly energy use was predicted with the least error, and daily energy was predicted with the most error

Another way to compare performance of models is to plot the errors from one model versus the errors from another model: for each building, the error from one model determines the location along the x-axis and the error from another model determines the location along the y-axis. 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 Figure 2, the nRMSE for the proprietary models is compared to that for the LBNL model, for each duration of training period and each predicted quantity.

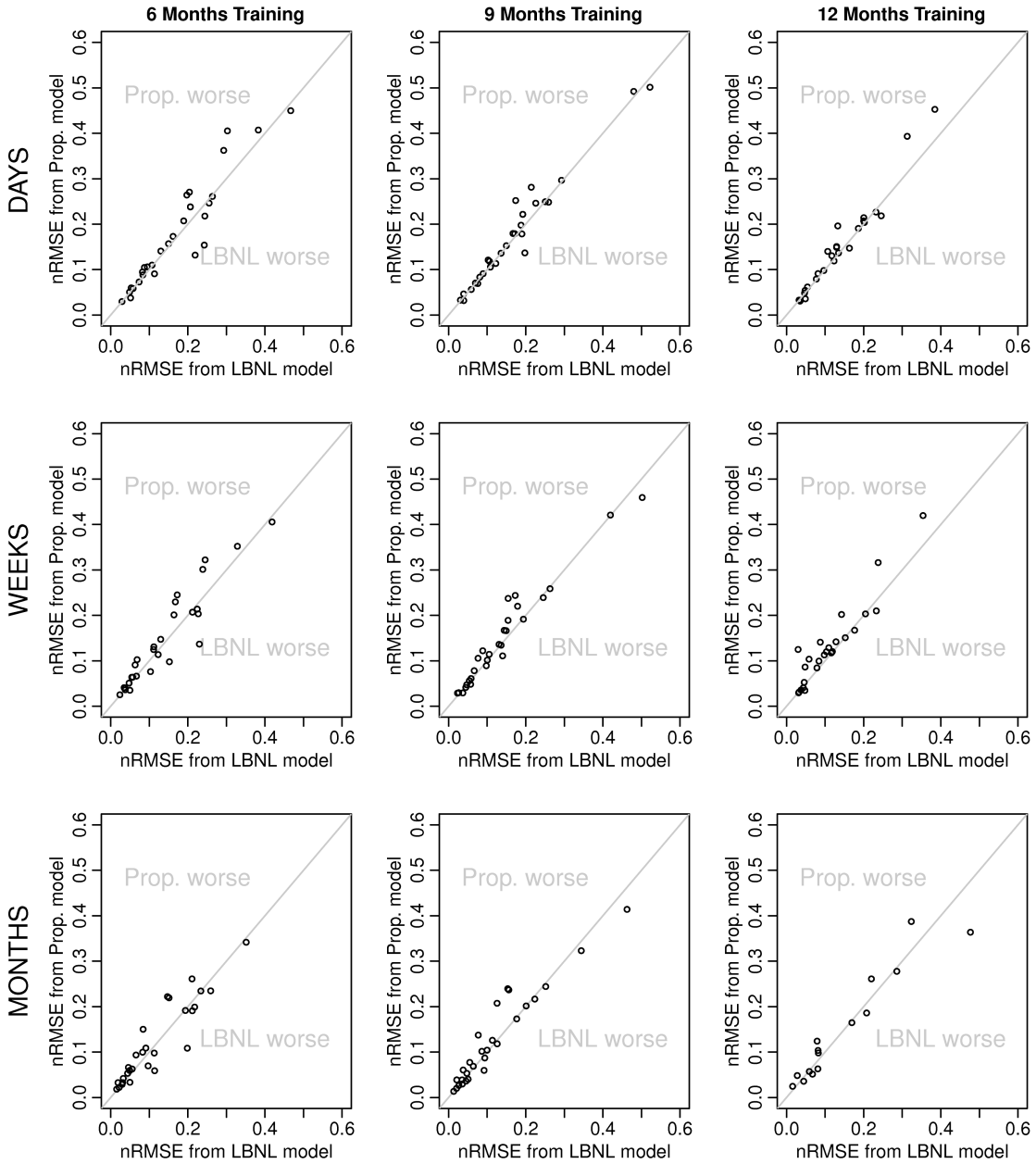

Figure 2: nRMSE of the Proprietary model vs. that of the LBNL model, for each building. Results are shown for daily, weekly, and monthly energy use predictions (rows) for 6-month, 9-month, and 12-month training periods (columns).

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)	PROPR. (.035)
	LBNL (.032)	MW (.049)	LBNL (.041)
	DTT (.034)	CP (.052)	DTT (.046)
	PROPR. (.037)	LBNL (.055)	CP (.065)
Worst	CP (.051)	PROPR. (.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%
Proprietary model, 4-6%
Mean-week model, 3-7%
Change-point model, 5-7%

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, compliance with ASHRAE Guideline 14, and implications concerning the use of these models for whole-building M&V applications.

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 proprietary model, LBNL model, and day-time-temperature (DTT) models perform very similarly with respect 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. Like all of the models, they 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 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 proprietary and LBNL models 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 proprietary models, 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 proprietary 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 proprietary, 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, proprietary, 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, the LBNL, proprietary, 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 Model Compliance with ASHRAE Guideline 14

ASHRAE Guideline 14 [8] 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 notably restrictive, in that 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 the proprietary model 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, that is, data from the training period. As such, neither is amenable to independent validation in a strict sense: a modeler can always make post-facto adjustments to force compliance. Although LBNL did not have access to predictions from the proprietary model'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 proprietary model was no more biased than the other models. Although it was not possible to independently test whether the proprietary complied with these requirements, the study results indicate that in general, the it predicted energy use more accurately than the 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 proprietary model all comfortably met required threshold for most of the buildings in the test data set; all failed to meet the requirement for the most unpredictable buildings. For monthly energy use predictions, which Guideline 14's whole-building discussion is centered upon, the three top performing models far exceeded the 20% threshold requirement, with median values of 8-11%.

5. Conclusion

The methodology developed and applied in this work comprises the first step in establishing a general approach to evaluate the predictive accuracy of whole-building

baseline models, which is a critical component of the uncertainty in the measurement and verification (M&V) of gross, whole-building energy savings. This methodology relies upon aggregating interval meter data from a number of buildings into a testing data set, and applying cross validation to compare model predictions to metered building energy data. The test data set is divided into a block of model training data, and a block of prediction data that follows the training period in time. Model performance is then evaluated using a number of statistical metrics, of which two, normalized root mean squared error and median absolute relative total error, were judged most critical to the consideration of uncertainty in determining energy savings. (Other performance metrics included in the methodology, such as correlation coefficient may by be more useful for other analysis methods that depend on baseline models, such as anomaly detection.) By varying the training and prediction periods it is possible to quantify model performance relative to diverse pre- and post-measure periods, which influence the time required to quantify energy savings. By varying the unit of energy prediction, for example, daily, vs. weekly, vs. monthly total energy use, one can also judge how errors change with different 'reporting intervals' that might be used by owners or program mangers to track savings as an efficiency project progresses over time.

This is a general methodology that can be applied to proprietary or 'open' models, and to system-level and submeter data and models. The training and prediction time periods may be adjusted.

To demonstrate and validate this methodology, and gain insights regarding the performance of interval meter data, five models were evaluated against a set of testing data from 29 buildings, spanning a diversity of climates, sizes, and geographical locations. Since this data set was relatively small, and not fully representative, widely generalizable conclusions regarding model performance cannot be established. However, several findings relevant to questions of the time requirements and accuracy of whole-building M&V resulted from this demonstration, and suggest topics for deeper exploration in future research.

Error was reduced with more resolved models that account for time: The change-point models, which do not take time into account, were outperformed by each of the more sophisticated models that explicitly include time as an independent variable. This has important implications for whole-building M&V, as change-point models have historically served as an industry standard. The increasing availability of interval meter data, with associated time-stamp information therefore has the potential to improve the accuracy of whole-building M&V.

Traditional rules of thumb may be overly conservative for today's improved models: The IPMVP recommends that whole-building approaches be limited to cases where savings are greater than ten percent [2]. However, for the top three performing models in this study, the median absolute error in total metered energy use over all training periods and prediction horizons considered ranged from three to six percent, suggesting that

some models may be able to resolve whole-building savings of less than ten percent, particularly if pre-screening is used to target buildings that the baseline model characterizes with a high degree of fitness. It is possible that the buildings in the present study are more predictable than typical buildings; full investigation of this issue will require a larger dataset, ideally one that is statistically representative of a class of buildings (such as all commercial buildings in a given climate zone).

A longer baseline period does not guarantee lower error: Interestingly for M&V applications, the total error was smallest for a shorter training period and a longer prediction horizon. In contrast to naïve expectations, a longer training period does not necessarily lead to a model that makes better predictions: a building's energy behavior changes from month to month – hours of operation are altered, equipment is replaced, employees may be added or removed – so extending the training period to include data from eight or ten months ago may not improve the model and in fact may make it worse. When the only available data were monthly billing data, long training periods were needed in order to determine how energy consumption varies with temperature because it was necessary to include both hot and cold months; with interval data, a few hot days and cold days can be sufficient, and these may occur in a span of just a few months.

Aggregation of energy predictions into larger 'chunks' can reduce error: Not surprisingly, monthly energy totals were predicted with less error than daily or weekly energy totals, which is probably good news for M&V since savings are not typically reported or tracked on a daily or weekly basis.

6. Future Work

Building energy use changes with time for several reasons: operational changes such as operating hours and thermostat settings; equipment changes such as replacement of lights and office equipment; changes in occupancy or in occupant behavior; and external factors such as outdoor air temperature and humidity. Baseline model prediction errors therefore depend on both how well the model makes use of the explanatory variables available to it, and on how much variation is caused by factors not explicitly included in the model. Whole-building baseline models are typically based only outdoor temperature and humidity, so the effect of the other sources of variation of energy consumption is not captured. If a model does not make proper use of its input data, the model can be improved; but if the building's energy consumption varies due to factors that are not provided as inputs to the model, the result is error that no model improvement can fix. The present paper reports on the accuracy of several models when applied to one set of buildings, using outdoor air temperature as the only explanatory variable.

Our future work on quantifying the performance of baseline models will focus on (1) compiling larger, statistically representative datasets, and (2) using those data sets to explore the possibility of identifying types of buildings that are more predictable than

others; this would allow screening of buildings to identify candidates for energy savings measures whose effectiveness can be reliably quantified through whole-building M&V. We also hope to (3) develop a formal methodology to evaluate the suitability of models for M&V (as applied to a given set of buildings) and (4) develop a way to determine what performance criteria must be met in order to meet the needs of a given M&V incentive program – for instance, 6-month energy use must be predictable in at least 50% of buildings to within 5%, and in 85% of buildings to within 8%.

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

[2] 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.

[3] Granderson, J, Piette, MA, Ghatikar, G. Building energy information systems: User case studies. 2011. *Energy Efficiency* 4(1):17-30.

[4] 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.

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

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

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

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

[9] 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.

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

[11] 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: Additional Statistical Performance Metrics Included in the Baseline Performance Evaluation Methodology

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

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[‡].

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

[‡] 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.

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 A.3: $$\mathrm{nMAE} = \frac{\frac{\sum_{i=1}^{n}\left|E_i - \hat{E}_i\right|}{n}}{\frac{\sum_{i=1}^{n}\left(E_i\right)}{n}}$$

The *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 A.4: $$\mathrm{medRTE} = \mathrm{median}\left\{\frac{\left(E_i - \hat{E}_i\right)}{E_i}\right\}$$

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.

Appendix B: Detailed Results, Statistical Performance Metrics

Detailed statistical performance metrics for each of the baseline models are provided in Tables B1- B10. Tables B1-B3 summarize the median model performance for daily energy use predictions, Tables B4-B6 correspond to weekly energy predictions, and Tables B7-B9 correspond to monthly energy predictions. Table A10 summarizes the relative total error for each model, training period, and prediction horizon.

Table B1. Model performance over the entire data set, daily energy predictions, 6-month training period, 10-month prediction horizon. Several goodness-of-fit metrics are shown, as well as quantiles of relative error over the entire prediction period.

Model	r	nRMSE	relMAE	relTotErr	2.50%	10%	50%	90%	97.50%
PROPR.	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
MW	0.483	0.204	0.162	0.009	-0.324	-0.174	0.008	0.174	0.297

Table B2. Median model performance over entire data set, daily energy predictions, 9-month training period, 7-month prediction horizon. Several goodness-of-fit metrics are shown, as well as quantiles of relative error over the entire prediction period.

Model	r	nRMSE	relMAE	relTotErr	2.50%	10%	50%	90%	97.50%
PROPR.	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 B3. Median model performance over entire data set, daily energy predictions, 12-month training period, 4-month prediction horizon. Several goodness-of-fit metrics are shown, as well as quantiles of relative error over the entire prediction period.

Model	r	nRMSE	relMAE	relTotErr	2.50%	10%	50%	90%	97.50%
PROPR.	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 B4. Median model performance over entire data set, weekly energy predictions, 6-month training period, 10-month prediction horizon. Several goodness-of-fit metrics are shown, as well as quantiles of relative error over the entire prediction period.

Model	r	nRMSE	relMAE	relTotErr	2.50%	10%	50%	90%	97.50%
PROPR.	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 B5. Median model performance over entire data set, weekly energy predictions, 9-month training period, 7-month prediction horizon. Several goodness-of-fit metrics are shown, as well as quantiles of relative error over the entire prediction period.

Model	r	nRMSE	relMAE	relTotErr	2.50%	10%	50%	90%	97.50%
PROPR.	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 B6. Median model performance over entire data set, weekly energy predictions, 12-month training period, 4-month prediction horizon. Several goodness-of-fit metrics are shown, as well as quantiles of relative error over the entire prediction period.

Model	r	nRMSE	relMAE	relTotErr	2.50%	10%	50%	90%	97.50%
PROPR.	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 B7. Median model performance over entire data set, monthly energy predictions, 6-month training period, 10-month prediction horizon. Several goodness-of-fit metrics are shown, as well as quantiles of relative error over the entire prediction period.

Model	r	nRMSE	relMAE	relTotErr	2.50%	10%	50%	90%	97.50%
PROPR.	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 B8. Median model performance over entire data set, monthly energy predictions, 9-month training period, 7-month prediction horizon. Several goodness-of-fit metrics are shown, as well as quantiles of relative error over the entire prediction period.

Model	r	nRMSE	relMAE	relTotErr	2.50%	10%	50%	90%	97.50%
PROPR.	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 B9. Median model performance over entire data set, monthly energy predictions, 12-month training period, 4-month prediction horizon. Several goodness-of-fit metrics are shown, as well as quantiles of relative error over the entire prediction period.

Model	r	nRMSE	relMAE	relTotErr	2.50%	10%	50%	90%	97.50%
PROPR.	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 B10. Median Relative Total Error for each model, predicted quantity, and training period. Several goodness-of-fit metrics are shown, as well as quantiles of relative error over the entire prediction 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)	PROPR. (-.008)
	MW (.009)	MW (-.012)	DTT (-.010)
	PROPR. (.011)	LBNL (.013)	MW (-.011)
	DTT (.012)	CP (-.013)	LBNL (.013)
Worst	CP (.016)	PROPR. (.026)	CP (.036)

Appendix C: Building Characteristics

Table C-1 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 C-1. Floor area and commercial type for buildings included in the study

Location	Area (sf)	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

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.