ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

Functional Testing Protocols for Commercial Building Efficiency Baseline Modeling Software

David Jump

Quantum Energy Services and Technologies, Inc. (QuEST)
2001 Addison St., Suite 300
Berkeley, CA 94704

Phillip N. Price, Jessica Granderson and Michael Sohn

Lawrence Berkeley National Laboratory
Environmental Energy Technologies Division
Berkeley, CA 94720

September 2013

This work was supported by Pacific Gas and Electric Company (PG&E) and by the Assistant Secretary for Energy Efficiency and Renewable Energy, Office of Building Technology, State and Community Programs, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231.

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.

Functional Testing Protocols for Commercial Building Efficiency Baseline Modeling Software

ET Project Number: ET 12PGE1311

Project Manager: **Leo Carillo**
Pacific Gas and Electric Company

Prepared By:

David Jump, Ph.D., P.E.
Quantum Energy Services & Technologies, Inc. (QuEST)
2001 Addison St. Suite 300
Berkeley, CA 94704

Phillip N. Price, Ph.D.
Jessica Granderson, Ph.D.
Michael Sohn, Ph.D.
Lawrence Berkeley National Laboratory (LBNL)
1 Cyclotron Road
Berkeley, CA 94720

Issued: **September 6, 2013**

Acknowledgements

Pacific Gas and Electric Company's Emerging Technologies Program is responsible for this project. It was developed as part of Pacific Gas and Electric Company's Emerging Technology – Technology Development Support program under internal project number ET12PGE1311. Quantum Energy Services & Technologies, Inc. (QuEST), with assistance from Lawrence Berkeley National Laboratory (LBNL), developed this protocol for Pacific Gas and Electric Company and overall guidance and management was provided by Leo Carrillo. For more information on this project, contact *lmcz*@pge.com.

The authors also wish to acknowledge all others who assisted this project, including Portland Energy Conservation Inc., Dr. Agami Reddy, and PGE's Mananya Chansanchai, Mangesh Basarkar, and Ken Gillespie, as well as the members of our the Technical Advisory Group that PG&E organized for this project. Special thanks to Gavin Hastings, Arizona Public Service Co; Glenda Towns, Southern California Gas Co.; and Graham Henderson, BC Hydro Inc. for their participation in a utility focus group on user requirements for the test protocols.

Legal Notice

This report was prepared for Pacific Gas and Electric Company for use by its employees and agents. Neither Pacific Gas and Electric Company nor any of its employees and agents:

(1) makes any written or oral warranty, expressed or implied, including, but not limited to those concerning merchantability or fitness for a particular purpose;

(2) assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, process, method, or policy contained herein; or

(3) represents that its use would not infringe any privately owned rights, including, but not limited to, patents, trademarks, or copyrights.

Table of Contents

Overview 1

Model Prequalifying Protocol 4

Purpose 4

Scope 5

Procedure 5

1. Establish Test Goals 6
2. Select Evaluation Metrics 8
3. Collect Building Data 10
4. Prepare Test Data Set 11
5. Calculate Predictions 11
6. Calculate Metrics and Evaluate Performance 13
7. Performance Benchmarking with Public Domain Models 17
8. Document Scorecard 20

Prequalifying Test Scorecard 21

Field Test Protocol 25

Purpose 25

Scope 25

Procedure 25

1. Establish Accuracy Requirements 27
2. Identify Customer and Collect Data 27
3. Prepare Test Data Set 28
4. Vendor Calculates Predictions 29
5. Calculate Metrics and Evaluate Performance 30
6. Document Scorecard 31

References 33

OVERVIEW

As advanced metering infrastructure including time-of-use and smart meter technology penetrates more and more buildings in utility service areas, an abundance of rich short-time interval data is becoming available. Software vendors are mining this data to provide energy management services to customers, establish energy baselines, and track performance over time.

The test protocols in this report were developed to guide testing of software products with baseline modeling and savings estimation functionality that may be used to estimate whole-building energy savings over a period of time. Savings measurement and verification (M&V) requires comparing the amount of energy actually used to the amount of energy the building would have used had energy efficiency measures not been implemented. The energy the building would have used cannot be measured; it must be estimated with a baseline model. The accuracy of the savings estimate is therefore directly dependent on the accuracy of the baseline model.

Baseline models are developed from energy use and independent variable data such as weather and operation schedule, and other information such as building size or use. The models range from simple regressions to more complex regressions of multiple parameters, neural networks, integrated moving averages, or innovative combinations of these techniques.

This document describes procedures for testing and validating proprietary baseline energy modeling software accuracy in predicting energy use over the period of interest, such as a month or a year. The procedures are designed according to the methodology used for public domain baselining software in another LBNL report that was (like the present report) prepared for Pacific Gas and Electric Company: "Commercial Building Energy Baseline Modeling Software: Performance Metrics and Method Testing with Open Source Models and Implications for Proprietary Software Testing Protocols" (referred to here as the "Model Analysis Report"). The test procedure focuses on the quality of the software's predictions rather than on the specific algorithms used to predict energy use. In this way the software vendor is not required to divulge or share proprietary information about how their software works, while enabling stakeholders to assess its performance.

Modeling accuracy and uncertainty are determined through comparisons of model predictions with actual building energy use for a data set of a representative sample of buildings. The protocol's fundamental procedures involve dividing up individual building energy use data into training and prediction periods, and using data from the training periods to create a model that predicts the energy use in the prediction periods. Performance metrics that characterize the accuracy of model predictions over time periods of interest are quantified. The distribution of

these metrics describes the uncertainty in the software's energy use predictions for the building population represented in the test data set. Please note that the terms 'performance metric, evaluation metric, and error metric are used synonymously in this protocol.

Public domain models are included and tested using the same procedure as the proprietary models. Their distribution of metrics provides an appropriate benchmark for evaluating the proprietary modeling methods with respect to estimating savings.

Two test protocols are provided: the first is a prequalifying protocol that enables prospective customers and stakeholders to assess model performance on the population of buildings with respect to savings estimation, and to compare the model's performance compared to a number of well-performing public domain models. To allow some flexibility in its application, the protocol describes two potential implementation paths: one where a prospective customer or its agent (a third party testing firm) obtains and runs the vendor's software and evaluates the results, and one where the vendor receives data from the third party, runs its software, and provides its predictions back to the third party for evaluation. Results of the prequalifying protocol are provided as a scorecard.

The second protocol is a field test where the proprietary model's performance is tested for a specific building. While a vendor's software may score well on the prequalifying protocol, the field test allows stakeholders to assess its performance for a specific building. Results of this test are pass or fail.

<u>These protocols are intended to inform the development of test protocols in the utility industry. While these protocols describe test procedures and provide options for their implementation, they are not intended to serve as the definitive protocols for PG&E or any other utility or stakeholder.</u>

There are several risks and issues that the prequalifying and field test protocols must navigate. These include:

- Protection of customer identifying information and energy use data. Utilities must safeguard customer data and information. The tests must be conducted with safeguards to assure that any data and information provided does not allow identification of customers by software vendors and/or any other unauthorized parties.
- Protection of software vendor's proprietary modeling algorithms. Software vendors invest a tremendous amount of funds and other resources into developing their products. The software tests must be conducted in a manner that protects competitors and other outsiders from obtaining their modeling and other software algorithms.
- Prevention of intervention in baseline modeling by the software vendor. Tests require the software products to be run without human intervention in order that the software's capabilities are tested, not that of any human-software combination. Engineers and other experts could potentially provide insight and interact with the

software to improve its energy predictions for the test, potentially creating an unfair advantage over other software participants.

- Establishment of appropriate 'blinds' to prevent influence on test results. Software vendors must not know the actual energy use that they are asked to predict using their modeling methods. This will prevent potential manipulation of results that make their predictions more accurate than their software is capable of accomplishing.

The protocols describe ways in which software evaluators may manage the potential risks these issues impose.

Future revisions of the testing protocols may include addition of specific sampling algorithms, reduction in number of building data sets, or revision of the recommended performance metrics, training and prediction period scenarios. Revisions to these protocols may be made based on test procedure viability, feedback from the software industry, or other issues.

This is a proposed testing procedure. It is proposed for use by energy efficiency program sponsors whose objective is to evaluate the baseline accuracy of software vendor's proprietary baseline and savings estimation methods. It is not a definitive test of model performance for several reasons, including the fact that the explanatory variable data made available to participating vendors may not be sufficient to achieve the higher levels of predictive accuracy for which that their tools are capable. The protocols are a work in progress. The authors seek feedback from such stakeholders as well as from the energy management software industry. For example, particular feedback of interest is whether vendors value software testing as a means to increase confidence in their products, and elect to develop software modules to facilitate third party testing.

MODEL PREQUALIFYING PROTOCOL

PURPOSE

The purpose of this testing protocol is to statistically determine how accurately proprietary energy baseline modeling software predicts energy usage for a population of buildings over a given time period, such as a month or year, and to quantify the uncertainty in those predictions. As documented in the Model Analysis Report, this protocol's methodology may be used to evaluate a proprietary model's performance for multiple use cases: besides energy use prediction, it may be used for performance tracking, demand response, and anomaly detection. However this protocol is designed to test a model's energy use prediction capability, as it is a key element of savings estimation.

Under this protocol, proprietary baseline modeling software is tested using a test data set that consists, at a minimum, of energy use data from multiple buildings of sufficient number to be representative of a population of buildings. Results of this test describe how accurately the software predicts energy use for the building population, whether the software's accuracy is acceptable for a specific application, and how well the software's accuracy compares to that of selected public domain models.

Software prediction accuracy and uncertainty are obtained from the distribution of errors between the model's predicted and actual energy use. Cumulative distribution function charts and tables are used to show the level of error achieved over the building population. They provide a means of assessing a model's predictive ability. In addition, the prediction errors from a model (for any time period of interest) may be charted and compared to prediction errors from a public domain model as a benchmark to determine which model achieved better performance.

This protocol is written to provide flexibility in its application, by describing alternatives and options to the test procedures whenever possible.

Please note the distinction between "test sponsors," who are utilities and energy efficiency program administrators entrusted with safeguarding customer information, including energy use data, and "test administrators" who may be utilities, energy efficiency program administrators, their third party representatives, or building owners, who develop the specific procedures and requirements for testing software vendor energy models.

SCOPE

The scope of this testing protocol includes any software that uses measured whole-building energy use and other data such as corresponding weather data, building characteristics, or operational schedules as deemed relevant by test administrators to predict whole-building baseline energy use, and ultimately to estimate savings. This protocol tests the accuracy of the baseline model predictions without requiring review of modeling algorithms or source code, thereby eliminating the need for in-depth technical assessment of model algorithms while also protecting a software vendor's proprietary algorithms and programming.

PROCEDURE

The baselining software test procedure is outlined in Figure 1. Each step in the procedure is described below. For some steps, alternative procedures and options that allow test administrators flexibility in adapting the test to their circumstances are described.

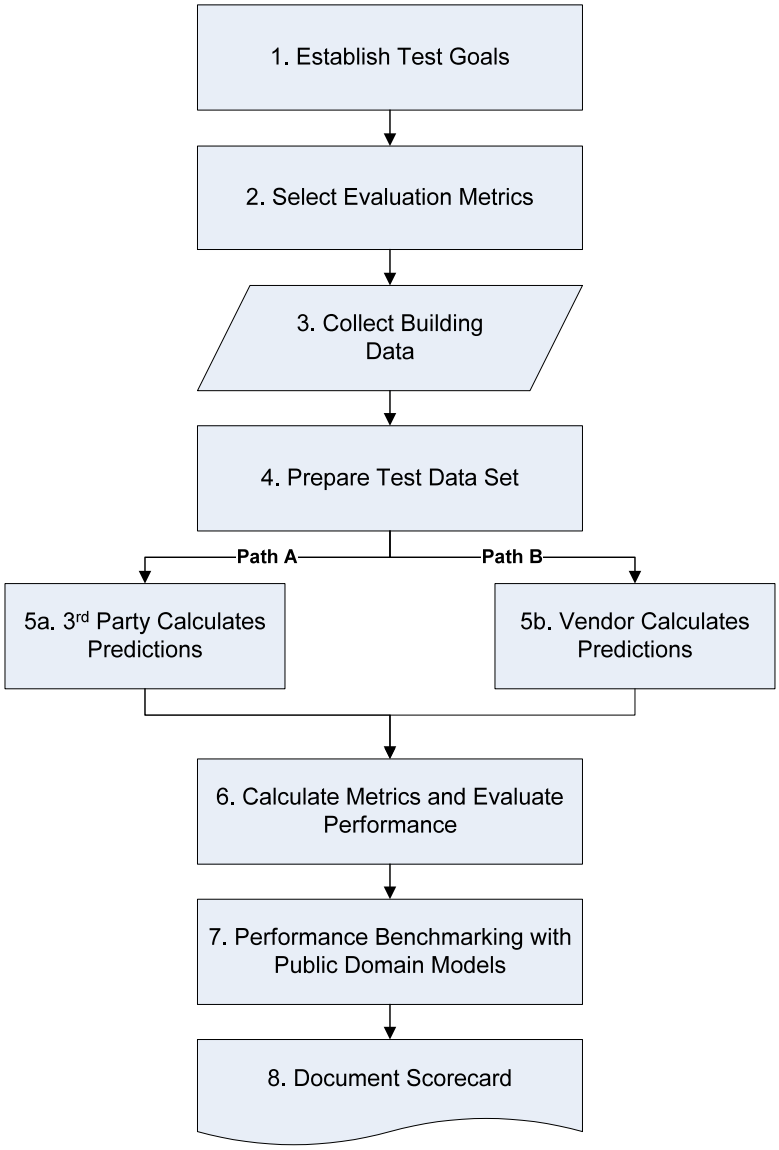


FIGURE 1. MODEL PREQUALIFYING TEST PROCEDURE.

1. ESTABLISH TEST GOALS

This test methodology was developed to enable evaluation of baseline energy prediction of proprietary software models both in terms of overall robustness and in comparison with other candidate models. Results of this evaluation inform stakeholders about the model's accuracy in predicting energy use and estimating savings over a period of time, such as a month or a year. The first step in conducting the test is for test administrators to establish the goals of the exercise. Establishing the goals first will help identify the proper performance metrics to be quantified. The goals may be established by answering some key questions, such as:

1. *What is the period of time over which baseline energy use will be estimated? A related question may be how often must savings be estimated after EEMs have been installed?*

Most utility incentives are paid based on an annual estimate of savings, therefore the accuracy in which a software model predicts annual energy use and estimates savings is of particular interest. In some cases, program administrators may require energy use and savings be estimated over more frequent intervals, such as a month, or a quarter (three months). The number of error metrics and training/prediction period durations may be adapted to accommodate assessment of model energy predictions over shorter time intervals.

2. *What is the level of risk, in terms of savings uncertainty, tolerable for each project?*

Risks that a project does not achieve savings come from many sources. One of those risks is using a savings calculation method that does not yield an estimate of uncertainty. This test can inform the level of risk associated with the amount of uncertainty in the software's calculation of savings. As a rule, the lower the uncertainty, the less risk, however no project can yield a savings estimate without uncertainty. Stakeholders should establish the level of uncertainty associated with savings estimation that is acceptable for their projects and programs. How this uncertainty level is established is beyond the scope of this protocol, however a simple example will be used to illustrate how to apply an uncertainty criterion in this protocol, as described below.

> For illustrative purposes only, the specific criterion selected was developed by considering a typical retro-commissioning project in a medium-sized commercial office building. For such a project, 15% of the building's annual electric energy use is expected to be saved (Mills et al., 2004). In this case, project sponsors have determined that the maximum allowable uncertainty in the savings shall be no more than half, or 7.5%, of the estimated savings, at a 90% confidence level. Project sponsors desire to have high confidence that the savings is within this precision level, hence the 90% confidence requirement (ASHRAE 2002, p. 102). Later, this criterion will be used to assess a model's predictive accuracy for projects of this nature.

It is important to note that the savings uncertainty level is different on a project-by-project basis as compared to a portfolio of projects. As documented in the Model Analysis Report, the uncertainty decreases as the population of projects in the portfolio increases. This means that while the savings uncertainty may be too high for any particular building in the portfolio, at the portfolio level, the uncertainty may be acceptable.

3. *Will the models be used for tracking a building's energy performance over time? If so, how frequently?*

Tracking a building's energy performance provides useful information to building operators about the efficient operation of building systems, and whether the building is on track to

achieve its goals. Actual building energy use may be compared with efficient energy use as predicted by a model. How accurate the model is over shorter time periods, such as an hour, a day, or a week becomes important to assess for this purpose.

2. Select Evaluation Metrics

Once the goals of the test are established, the set of evaluation metrics may be selected. These metrics are calculated using each vendor's prediction and actual measured whole-building energy use over the period of interest.

Absolute Percent Bias Error:

For determining the baseline model's accuracy in annual energy use predictions, the *absolute percent bias error* metric should be calculated:

$$APBE = \left| \frac{\sum_{i=1}^{N} \hat{E}_i - \sum_{i=1}^{N} E_i}{\sum_{i=1}^{N} E_i} \right| \times 100\% \qquad (1)$$

Where: $\hat{E}_i$ is the model's predicted energy use for the analysis time interval *i*, E_i is the interval's measured energy use, and *N* is the total number of predicted and measured points in the year. Note that *N* changes with the analysis time interval, such as 15-minute, hourly, daily, or monthly.

For determining the baseline model's accuracy over shorter prediction periods, the *mean absolute percent error (MAPE)* metric for a three-month quarter or on a monthly basis may be calculated. In this case an annual average of the quarterly or monthly error metric is calculated:

Quarterly MAPE:

$$MAPE_{quarter} = \frac{\sum_{q=1}^{4} \left| \frac{\hat{E}_q - E_q}{E_q} \right| \times 100\%}{4} \qquad (2)$$

Where: $\hat{E}_q$ is the model's quarterly energy use prediction, E_q is the quarterly measured energy use, and *q* is the quarter index.

Monthly MAPE:

$$MAPE_{month} = \frac{\sum_{m=1}^{12} \left| \frac{\hat{E}_m - E_m}{E_m} \right| \times 100\%}{12} \quad (3)$$

Where: $\hat{E}_m$ is the model's monthly energy use prediction, E_m is the monthly measured energy use, and *m* is the month index.

The *normalized root-mean-squared error* n(RMSE) provides an indication of how much the model's predictions vary from the actual values for the time intervals of interest, which for performance tracking may be an hour or a day.

n(RMSE) on an hourly basis (assumes data is available in hourly intervals, or is obtained by summing sub-hourly time interval data and predictions):

$$n(RMSE) = \frac{\sqrt{\frac{\sum_{i=1}^{N} \left(E_i - \hat{E}_i\right)^2}{N}}}{\overline{E}} \quad (4)$$

Where $\overline{E} = \frac{\sum_{i=1}^{N} E_i}{N}$ is the average energy use per hour, and *N* = total hourly intervals in the prediction period.

n(RMSE) on a daily basis (assumes sub-daily data is summed to daily intervals first):

$E_d = \sum_{i=1}^{24} E_i$ for each day in prediction data set

$$n(RMSE) = \frac{\sqrt{\frac{\sum_{d=1}^{D} \left(E_d - \hat{E}_d\right)^2}{D}}}{\overline{E}} \quad (5)$$

Where $\overline{E} = \frac{\sum_{d=1}^{D} E_d}{D}$ is the average daily energy use, and D = total daily intervals in the prediction period.

3. Collect Building Data

The prequalification test requires at least two years of energy use and independent variable data for each individual building. The time interval of measurement may be as frequent as 15 minutes, as is common with electric energy data from time-of-use and smart meters, or hourly or daily intervals, as is common with natural gas smart meter data. Two years of data are required in order to set up training and prediction period scenarios, as described in the next section.

The number of buildings should be large in order to be a representative sample of the population of buildings where the software will be used. Using the statistical definition of a "Normal" (or "Gaussian") distribution, sample selection is determined by a few key factors: how precisely the sample should represent the population, the confidence associated with the desired precision level, and the population's variation (i.e. coefficient of variation of the standard deviation of the population, or CV(STD).) Appendix A of IPMVP (2012) includes calculations to determine the number of samples needed in order to estimate a quantity to a desired level of precision, depending on the intra-sample variation and the level of certainty required, and assuming the samples are drawn from a Normal distribution. For example, in order to have 90% confidence of estimating a mean value within 10%, if the individual samples have a coefficient of variation of 0.5 (i.e. 50%), approximately 70 samples are needed. The statistical distribution of bias errors in the public domain models was not a Normal distribution in the Model Analysis Report, which included 368 buildings. A more precise definition of sample size for this purpose requires further study. To be conservative, a representative sample population should exceed by at least 50% the number estimated for Normal distributions. This means a sample of about $(1 + 0.5) \times 70 = 105$ buildings should sufficiently represent the building population.

The buildings represented in the sample population should be selected at random, and not through self-selected volunteers, or buildings that have recently completed an energy efficiency project, or participated in an energy efficiency program. Selection from such convenient lists potentially introduces bias in the sample, which would render the sample non-representative of the population.

Independent variable data is required in order to develop energy models. This data includes parameters that are a significant driver of energy use in a building, and may include ambient weather, solar load, operation schedule, or building occupancy. The ambient weather data is typically the most accessible source of independent variable data. Depending on the source, it typically includes dry-bulb temperature, relative humidity, dew point temperature, barometric pressure, amount of rainfall, and wind speed and direction. While most models use the dry bulb temperature, many make use of the other available weather data. Reliable sources for solar data are less common, and collecting operation schedule or building occupancy data is generally unrealistic. It is important to note that while limiting the types of independent

variable data is advantageous for the test by creating a level playing field for testing all models, performance for individual models may suffer without key data streams they depend on.

Test administrators generally will collect and provide the independent variable data, however in the case of ambient weather, they may allow the software vendors to collect it by providing zip code or climate zone. Independent variable data must be collected for the same time period as the energy data.

Test sponsors may elect to provide other information to the software vendors, such as building size, number of occupants, or space use type. However as will be discussed in more detail below, test sponsors must take care not to divulge too much information that allows the customer to be identified, which not only violates customer security requirements, but also potentially creates an advantage for one vendor over another in the software test.

4. Prepare Test Data Set

Each vendor's energy use prediction software participating in the test requires a set of data to train its models and independent variable data for predicting energy use. These predictions are compared to the actual energy use through the use of performance metrics described above in Step 2. Different training and prediction period durations may be used in the test; we describe these by the notation: T:P, where T is the training period duration (in months) and P is the prediction period duration. Generally, we want to know how accurate a vendor's model is for a 12 month prediction period, but predictive accuracy for quarterly or monthly periods may also be of interest.

Test administrators may also desire to know whether a model can accurately predict energy use for a subsequent year based on shorter training periods, such as three and six months. A school of thought holds that the most recent building energy use patterns prior to a prediction period are the most likely to be representative of the prediction period. Therefore, shorter training periods may be used in the test and compared with results from longer training periods. For each T:P scenario, the concurrent ambient weather and other independent variable data must be provided.

Note that the number of model runs multiplies for each building in the data set, for each T:P scenario analyzed, and for each model included in the test. Standardized input and output file formats and file naming conventions are imperative, so that the large number of data sets may be processed in a reasonable amount of time, and troubleshooting programming errors is made easier.

5. Calculate Predictions

Preparation of the test data set is dependent on who will run the proprietary models. The choices are the vendors themselves, or an independent third party testing entity engaged by

the stakeholders and under non-disclosure agreements. This creates two pathways for obtaining the vendor's energy predictions. Under the third party pathway (Path A in Figure 1), vendors provide their compiled software (not source code) to the third party. The third party will run the vendor's model in batch mode and collect its predictions for further processing. The key attributes of this pathway are:

1. Vendor access to customer data is prevented, as no data is provided to the vendor. There is little need to take steps to conceal customer identifying information, including the energy data itself, as the third party is under agreement to not disclose this information
2. The vendor will provide its software and a license for its use, and possibly provide training on software use to the third party. However, the vendor's intellectual property is safeguarded, as the third party does not have access to code, only executable software, or is contractually bound not to disclose it. The third party must use an operating system that will support running the vendor's executable code. The third party evaluator may also sign a non-disclosure agreement with the vendor.
3. This path prevents manual intervention in the software. The models will be automatically run, with energy predictions resulting purely from the model's programmed algorithms. Intervention in baseline modeling by the vendor is not possible since the vendor is not running its software.
4. The third party evaluator controls the input and output data sets entirely. No additional data is used in the evaluation other than specified in the test procedures.

Under the vendor pathway (Path B in Figure 1), the vendor receives the data to train its model and predict energy use. The vendor provides its predictions back to the third party who then calculates and evaluates the error metrics. The key attributes of this pathway are:

1. Vendors operate their own software and provide predictions to the test administrator. This requires that they are provided energy and independent variable data. The test sponsor will determine what independent variable data, in addition to energy data, to make available for the vendors. For example, the test sponsor may make available weather station data or vendors may be required to collect this data independently. Customer-specific information such as zip code, street addresses, city, space use type, or building size may also be provided, but test sponsor requirements with regard to customer data privacy and confidentiality may limit the types of data available. In order to protect customer identifying information, the energy data may be transformed, although caution should be taken to ensure that any transformation does not compromise the quality of model test results.
2. Vendors are not required to provide their software, only its predictions of energy use. Therefore vendor's intellectual property is protected. Vendors must identify the software version being tested.

3. This test does not control for manual intervention in the software on the part of vendors. Ideally, the models should be automatically run, with energy predictions resulting purely from the models programmed algorithms. Under Path B there are no controls to assure automatic running of the vendor's models. However, due to the very large number of data sets required for testing, and potentially short turn-around times, manual intervention may not be practical.
4. Vendors will be provided multiple data sets for use in training their models and only independent variable data (not energy data) for the prediction period. Vendors are free to use additional information they deem appropriate in their analysis (such as weather data).
5. Example input data files in ASCII format will be provided. Explicit instructions and example files for vendors to return their energy use predictions will also be provided.

6. Calculate Metrics and Evaluate Performance

For each T:P scenario, each vendor's predicted energy use for each building in the data set is compared to the actual energy use using the selected evaluation metrics. For N buildings, N values of each error metric are calculated for each vendor model and each scenario. In addition, values of each error metric will be calculated for each public domain model's predictions for each scenario.

As an example, a prequalification test was developed where stakeholders desired information to decide between two proprietary vendor models. Their objective was to determine how accurate each proprietary model predicted annual energy use, using 12, 6, and 3 month training periods. In addition, they wanted to understand what improvement in energy use predictions the proprietary models had over a typical public domain model.

In this example, there are three T:P scenarios (3:12, 6:12, and 12:12), two proprietary models and one public domain model, and one error metric. For a data set consisting of 100 buildings, each T:P scenario requires 100 x 3 x 1 = 300 calculations, or a total of 900 calculations for all scenarios and models. The number of calculations increases with every additional error metric, scenario, and model evaluated.

Multiple metrics should be used in the evaluation to assess a model's predictive capability over a year, a three-month quarter, and a month. These include the *absolute percent bias error* (APBE) in annual predictions, the *quarterly mean absolute percent error* (quarterly MAPE), and *monthly mean absolute percent error* (monthly MAPE). If the models show good predictive accuracy over these time periods, stakeholders may elect to understand their accuracy on an hourly or daily basis using the hourly or daily n(RMSE).

In the ideal case, the public domain models should be run on the same data as the proprietary models. However to limit test costs, test sponsors may elect to use the results from the Model Analysis Report as a benchmark for comparison. Note there are several deficiencies in using these results for comparison: (1) the data set is likely not representative of the test sponsor's

building population; (2) the Model Analysis Report includes only the APBE and monthly MAPE evaluation metrics.

For each model and scenario, each metric's calculated values for the data set are charted as cumulative distribution functions. The values of the error metric are sorted from lowest to highest, and plotted against the percent of buildings achieving each value. An example for APBE for the Mean Week model is shown in Figure 2, and an example for monthly MAPE is shown in Figure 3. These charts may be read to see what each percentage of buildings achieved certain levels of the error metric. The more they shift to the left, the more accurate the model.

Tables of quantiles show the error metric results for specified percentiles of buildings. Figure 2 and Table 1 summarize the APBE distribution; Figure 3 and Table 2 summarize the distribution of montly MAPE.

Quantile tables are helpful to understand the uncertainty in using a particular model on a building represented by the dataset. Table 1 can be read to say a randomly selected building from the general population of buildings represented by the test data set has a 10% chance that the Mean Week model will predict energy use within 0.68%, a 20% chance the prediction is within 1.58%, and so on.

For the whole-building savings estimation use case, the uncertainty in the savings estimation is directly proportional to the uncertainty in the baseline model's energy use prediction. In general, a large savings compared to the uncertainty is desired. Since the distribution of bias errors is not a normal distribution, the assessment of expected savings could be compared to some high percentile, such as the 90th percentile. For the bias error shown in Table 1, this means that there is a 90% chance that the Mean Week model is wrong by less than 19% high or 19% low. This implies that the evaluation criteria for assessing model performance must consider the magnitude of savings, and the stakeholder's level of tolerance for savings uncertainty.

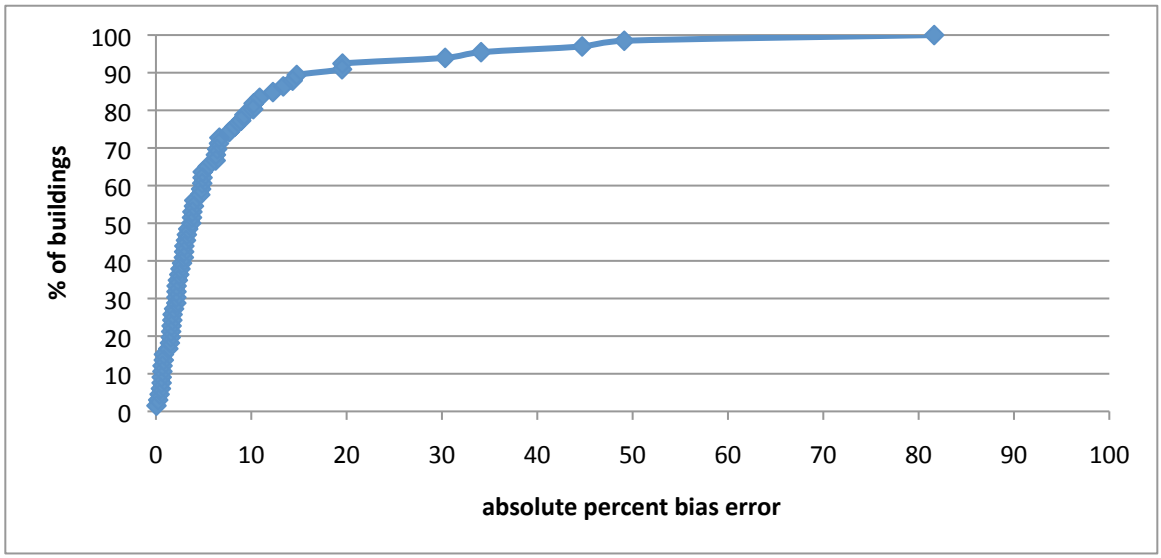


FIGURE 2. CUMULATIVE DISTRIBUTION FUNCTION OF APBE IN 12:12 PREDICTIONS OF THE MEAN WEEK MODEL.

TABLE 1. QUANTILES AND OF APBE FOR THE 100 BUILDINGS IN THE DATASET FOR THE MEAN WEEK MODEL. RESULTS FOR 12 MONTH TRAINING AND 12 MONTH PREDICTION PERIODS.

% buildings	10	20	50	80	90	mean
abs % bias error	0.68	1.58	3.75	10.22	19.04	8.12

For the illustrative example described earlier, project sponsors established a tolerance for savings uncertainty that was half the expected 15% savings typical of a retro-commissioning project, or 7.5%. If the Mean Week model was evaluated under this criterion, we would find it to be unacceptable: at a 90% confidence level, it was accurate only to within 19%. From Figure 2, The Mean Week model generates this level of error for approximately 65% of buildings. This percentile of buildings may be used in the scorecard to rank model performance. Note that similar criteria may be developed for monthly MAPE, quarterly MAPE or other error metrics.

For this dataset, the 7.5% criterion is very stringent, and it is doubtful that any other model would meet this performance level. To address this issue, project sponsors may elect to screen the candidate building population by eliminating buildings that are difficult to predict using common baseline models. Screening criteria are discussed in the Model Analysis Report.

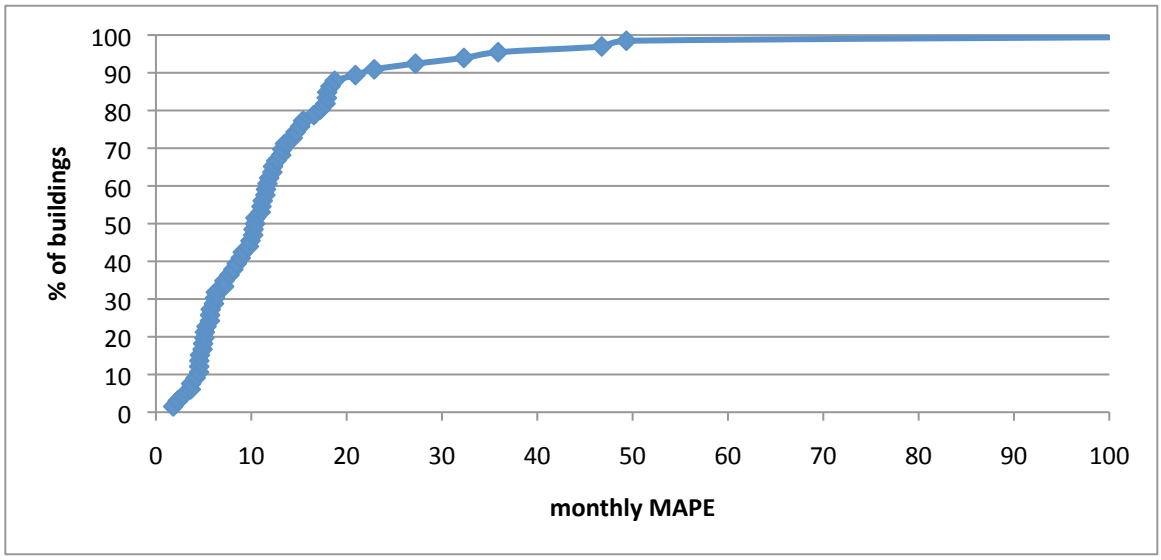


FIGURE 3. CUMULATIVE DISTRIBUTION FUNCTION OF MONTHLY MAPE IN 12:12 PREDICTIONS OF THE MEAN WEEK MODEL

TABLE 2. QUANTILES AND MEAN OF MONTHLY MAPE FOR THE 100 BUILDINGS IN THE DATASET FOR THE MEAN WEEK MODEL. RESULTS FOR 12 MONTH TRAINING AND 12 MONTH PREDICTION PERIODS.

% buildings	10	20	50	80	90	mean
monthly MAPE	4.53	5.11	10.42	17.43	22.70	14.10

Comparison charts directly compare one model's performance with that of another model, as shown in Figure 4. These charts show the values of the error metric calculated by each model for each building, and clearly show when one model is more accurate than the other. The results show that for this dataset, both models perform well at the lower APBE, while the Time-of-Week-and-Temperature model has a slight advantage for buildings with higher APBE.

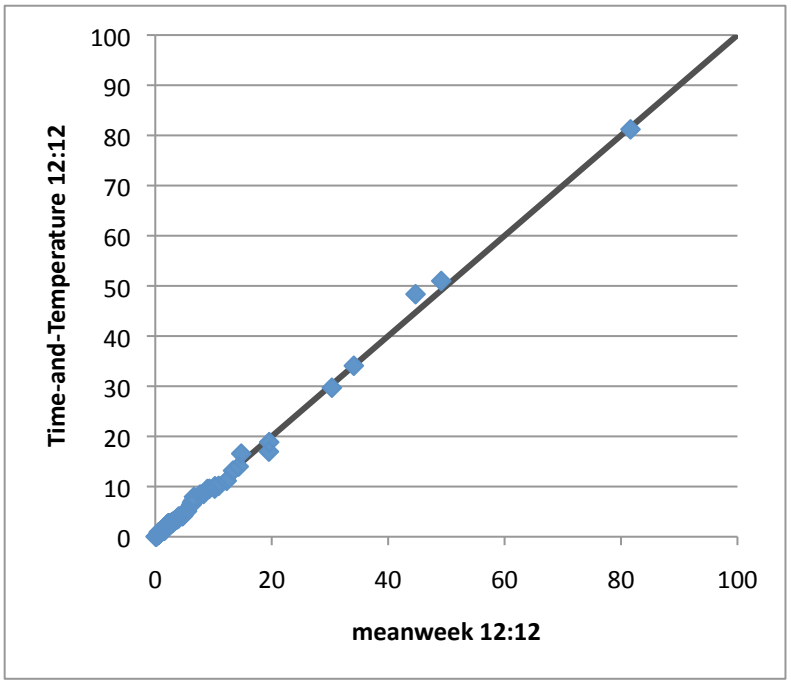


FIGURE 4. COMPARISON CHART OF APBE RESULTS FOR MEAN WEEK AND TIME-OF-WEEK-AND-TEMPERATURE MODELS. 12 MONTH TRAINING, 12 MONTH PREDICTION PERIODS.

7. PERFORMANCE BENCHMARKING WITH PUBLIC DOMAIN MODELS

If a test sponsor may find it useful to compare proprietary model performance with that of available public domain models, the third party must run selected public domain models. One or more selected public domain models must be run in parallel with the proprietary models using the same test data sets and the same error metrics are calculated. Brief descriptions of some public domain baselining models are provided below. Most of them rely on ambient dry-bulb temperature and time of week for independent variables. The quality of their predictions varies according to their modeling approach; some models may generate poor energy use predictions while others are very accurate. It is recommended to select at least two public domain models for the test in order that a wide range of prediction accuracy may be anticipated. The selected modules should be set up to accept the same input data sets as the proprietary models, and provide their prediction outputs in the same format as well.

COOLING DEGREE-DAYS AND HEATING-DEGREE-DAYS MODEL

The Cooling-Degree-Days and Heating-Degree-Days (CDD-HDD) model was originally developed for analyzing monthly billing data. For each month the number of heating and cooling degree-days is calculated, and linear regression is performed to predict monthly energy usage as a function of CDD and HDD. CDD and HDD were defined with base temperatures of 55 F and 65 F, respectively.

With m identifying the month, the model can be expressed as:

$$\hat{Q}_m = \beta_0 + \beta_c CDD_m + \beta_H HDD_m \tag{6}$$

Cooling- and Heating Degree-Days models provide a historical benchmark of comparison, however this comparison can only be on an annual basis.

MEAN WEEK MODEL

The predicted energy use for each time of day, for each day of the week, is equal to the mean for that time of day and day of week in the training dataset. For example, the prediction for every Monday at 1:00 PM is the average load for all of the Mondays at 1:00 PM in the training data.

As the Mean Week model has no weather dependence, it is expected to be less accurate for buildings with high weather sensitivity.

DAY-TIME-TEMPERATURE REGRESSION

In the Day-Time-Temperature model the predicted load is a sum of several terms: (1) a "day effect" that allows each day of the week to have a different predicted load; (2) an "hour effect" that allows each hour of the day to have a different predicted load; (3) an effect of temperature that is 0 for temperatures above 50 F and is linear in temperature for temperatures below 50 F; and (4) an effect of temperature that is 0 for temperatures below 65 F and is linear in temperature for temperatures above 65 F.

We define the following: i identifies the data point, day_i and $hour_i$ are the day and hour of that data point; T_{Ci} = 0 if the temperature T_i exceeds 50 F and is equal to 50 F -T if T_i < 50 F; T_{Hi} = 0 if T_i < 65 F and is equal to T_i - 65 F if T_i > 65 F.

With these definitions, the model can be written as

$$\hat{Q}_i = \beta_{day_i} + \beta_{hour_i} + \beta_c T_{c_i} + \beta_H T_{H_i} \tag{7}$$

The model is fit with ordinary regression.

The Day-Time-Temperature model considers time of week as well as ambient temperature, and is expected to be more accurate than other models that consider only one independent variable, such as the mean week model or the ASHRAE change-point models described below.

ASHRAE CHANGE-POINT MODELS

The American Society of Heating, Refrigeration and Air-conditioning Engineer's (ASHRAE) Research Project 1050 (Kissock et al., 2002) developed multi-parameter piecewise linear regression models, collectively known as change-point models. The number of model types tested and documented in the project includes:

1-parameter model (energy use is a constant): $E = C$ (8)

2-parameter model (simple linear regression): $E = C + B_1T$ (9)

3-parameter heating and cooling models:

Heating (downward slope until change-point at B_2, then flat): $E = C + B_1(B_2 - T)^+$ (10)

Cooling (flat until change-point at B_2, then upward slope): $E = C + B_1(T - B_2)^+$ (11)

4-parameter heating and cooling models:

Heating (downward slope, change in slope at B_3), : $E = C + B_1(B_3 - T)^+ - B_2(T - B_3)^+$ (12)

Cooling (upward slope, change in slope at B_3): $E = C - B_1(B_3 - T)^+ + B_2(T - B_3)^+$ (13)

In addition, there is a 5-parameter change point model for buildings with one energy source for both heating and cooling. It will not be referenced here. RP 1050 produced FORTRAN code that can be compiled and used in a batch process. Both the research report and code are available for a fee at www.ashrae.org.

The ASHRAE Change-Point models consider only ambient temperature, and are expected to be less accurate for buildings without high temperature sensitivity, or with energy use heavily dependent on operation schedules.

The model described in Equation 7 can be thought of a variant of the ASHRAE five-parameter change-point model. Unlike an ASHRAE five-parameter change-point model, it has fixed points for the temperature slopes (at 50 F and 65 F), and it adds time-of-day and day-of-week variation.

TIME-OF-WEEK-AND-TEMPERATURE REGRESSION

In the Time-of-Week-and-Temperature model, the predicted load is a sum of two terms: (1) a "time of week effect" that allows each time of the week to have a different predicted load from the others, and (2) a piecewise-continuous effect of temperature. The temperature effect is estimated separately for periods of the day with high and low load, to capture different temperature slopes for occupied and unoccupied building modes. The model is described in Mathieu et al. (2011). For this study the separation of hours of the week into "occupied" and "unoccupied" categories was automated: For each day of the week, the 10th and 90th percentile of the load were calculated; call these L10 and L90. The first time of that day at which the load usually exceeds the L10 + 0.1*(L90-L10) is defined as the start of the "occupied" period for that day of the week, and the first time at which it usually falls below that level later in the day is defined as the end of the "occupied" period for that day of the week.

The Time-of-Week-and-Temperature model considers time of week as well as ambient temperature, and is expected to be more accurate for predicting the load at a given time than

other models that consider only one independent variable, such as the mean week model or the ASHRAE change-point models. However, when aggregated over a long time period the model may not have a substantial advantage over others.

8. Document Scorecard

As described earlier, results of the calculation of the error metrics may be presented in ways to provide evaluators with convenient access to the information. A scorecard may be developed to describe the test conditions, the evaluation criteria, and results of the prequalification test. The results will show how the models perform on an absolute basis for the data set, how the models perform for a specific example project, and how models perform relative to public domain models. An example scorecard is provided in the next section. It shows the distributions of selected error metrics, quantile and mean tables of the selected metrics, and comparison charts between the vendor model and a public domain model (mean week).

This scorecard format may be adapted to include any other error metric desired, additional proprietary or public domain models, or T:P scenarios. When the data set is screened, the same error metrics, example project requirements, and T:P scenarios for each model may be reported in the same format.

PREQUALIFYING TEST SCORECARD

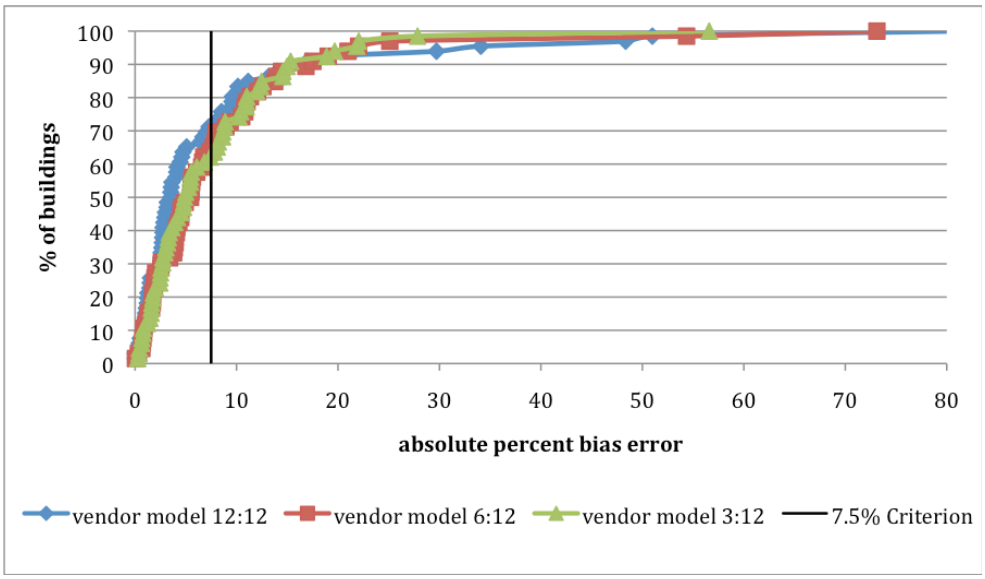


FIGURE 5. DISTRIBUTION OF APBE FOR 12:12, 6:12, AND 3:12 SCENARIOS, VENDOR MODEL. EXAMPLE PROJECT REQUIREMENT OF 7.5% APBE SHOWN.

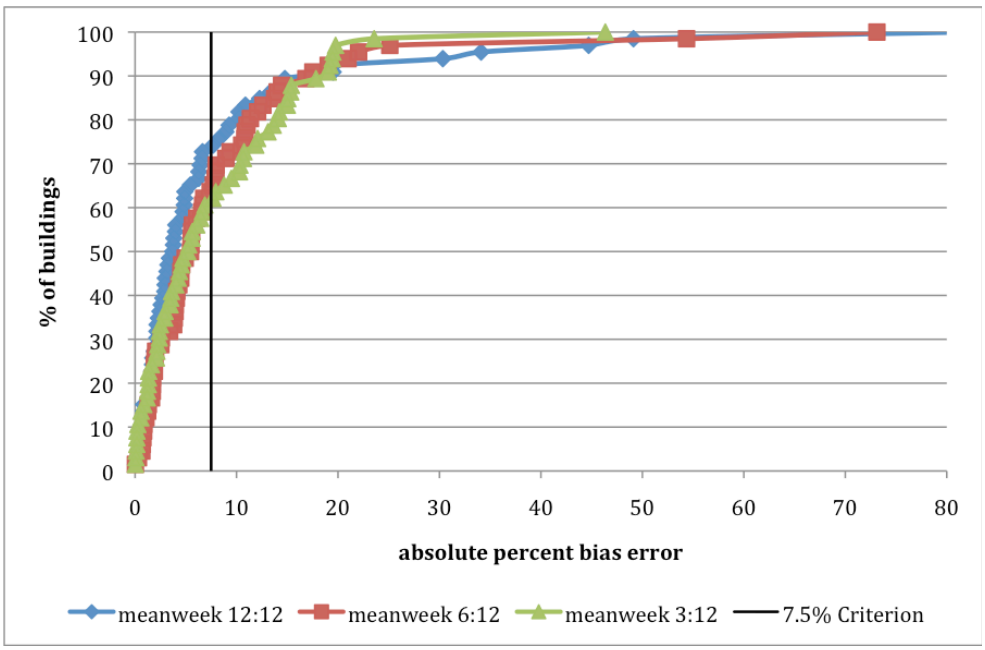


FIGURE 6. DISTRIBUTION OF APBE FOR 12:12, 6:12, AND 3:12 SCENARIOS, MEAN WEEK MODEL. EXAMPLE PROJECT REQUIREMENT OF 7.5% APBE SHOWN.

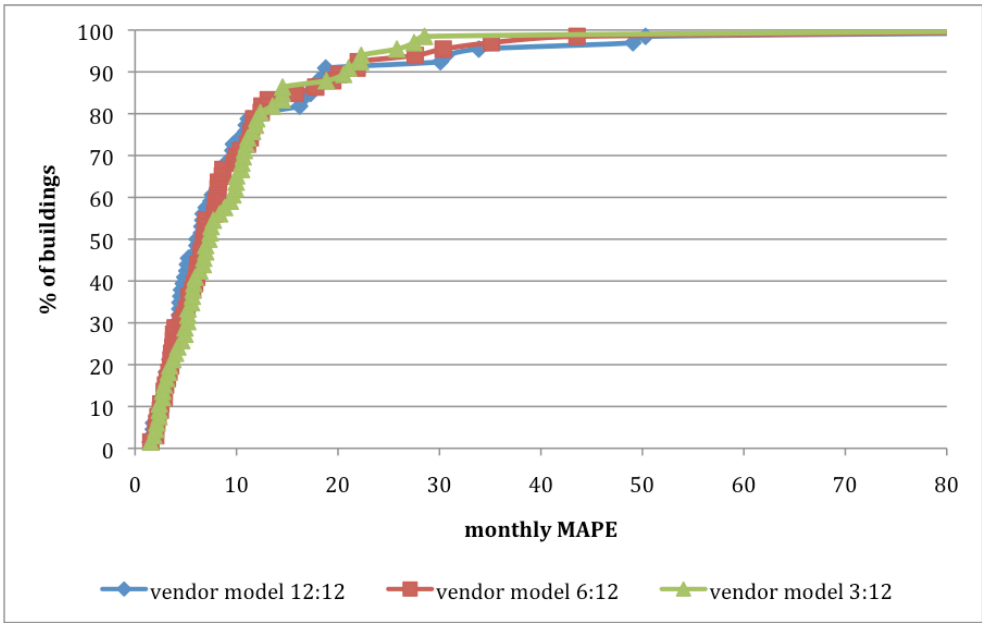


FIGURE 7. DISTRIBUTION OF MONTHLY MAPE FOR 12:12, 6:12, AND 3:12 SCENARIOS, VENDOR MODEL.

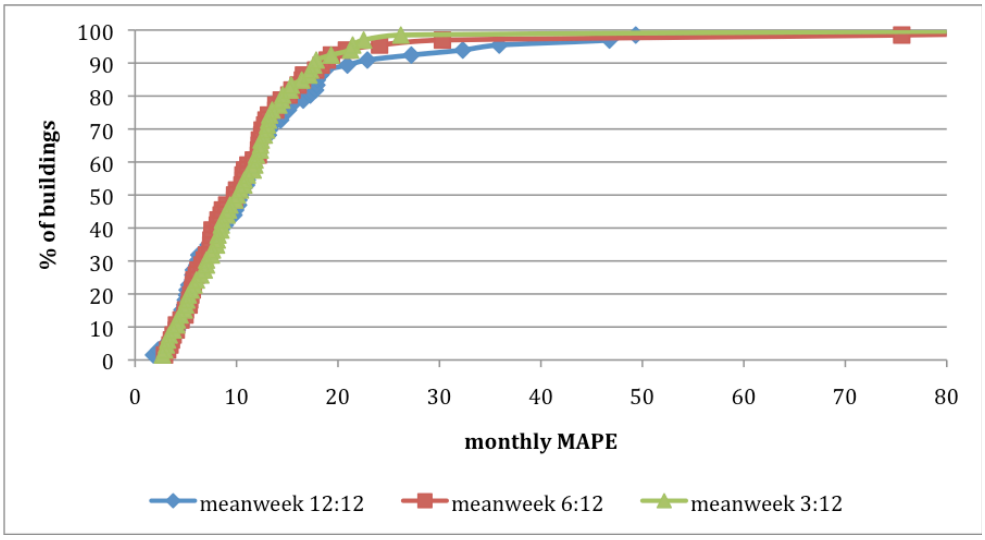


FIGURE 8. DISTRIBUTION OF MONTHLY MAPE FOR 12:12, 6:12, AND 3:12 SCENARIOS, MEAN WEEK MODEL.

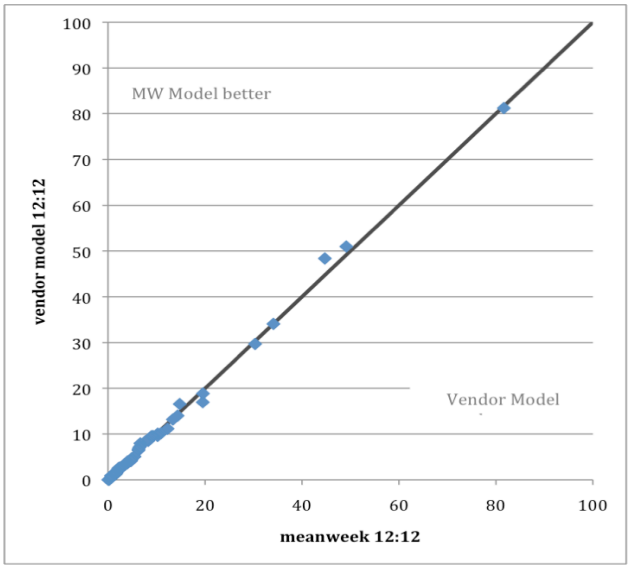


FIGURE 9. COMPARISON OF VENDOR AND MEAN WEEK MODEL RESULTS FOR APBE. 12:12 SCENARIO.

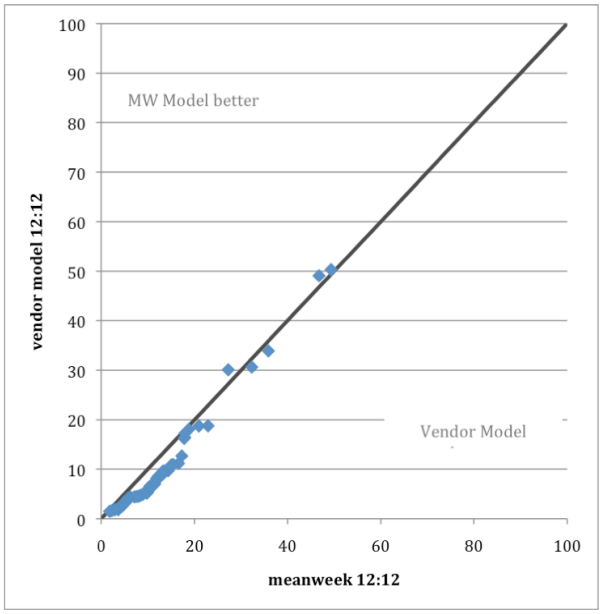


FIGURE 10. COMPARISON OF VENDOR AND MEAN WEEK MODEL RESULTS FOR MONTHLY MAPE. 12:12 SCENARIO.

TABLE 3. QUANTILES AND MEAN OF APBE FOR 3 T:P SCENARIOS, VENDOR AND MEAN WEEK MODELS. ALSO SHOWN ARE THE % OF BUILDINGS THAT MEET THE EXAMPLE 7.5% APBE REQUIREMENT.

Model	T:P Scenario	% buildings					mean	% buildings at 7.5% APBE
		10	20	50	80	90		
Vendor Model	12:12	0.90	1.25	3.52	9.73	16.92	8.14	71.70
	6:12	1.19	1.87	4.32	9.76	14.07	7.74	68.17
	3:12	0.87	1.78	4.94	11.31	15.30	7.60	62.48
Mean Week Model	12:12	0.68	1.58	3.75	10.22	19.04	8.12	74.01
	6:12	0.90	1.73	5.49	11.58	17.46	8.47	63.78
	3:12	0.31	1.31	5.24	14.14	18.92	7.75	61.80

TABLE 4. QUANTILES AND MEAN OF APBE FOR 3 T:P SCENARIOS, VENDOR AND MEAN WEEK MODELS. ALSO SHOWN ARE THE % OF BUILDINGS THAT MEET THE EXAMPLE 7.5% APBE REQUIREMENT.

Model	T:P Scenario	% buildings					mean
		10	20	50	80	90	
Vendor Model	12:12	2.58	3.42	6.32	13.74	18.79	11.28
	6:12	2.57	3.53	6.85	12.46	21.70	10.86
	3:12	2.42	3.63	7.34	12.71	21.00	10.44
Mean Week Model	12:12	4.53	5.11	10.42	17.43	22.70	14.10
	6:12	4.13	5.59	9.91	15.25	18.89	12.82
	3:12	4.11	5.44	10.23	14.77	17.83	11.89

FIELD TEST PROTOCOL

PURPOSE

The purpose of the Field Test Protocol is to determine how accurately a proprietary energy baseline modeling software predicts actual measured energy usage for a particular building over a given time period, and to quantify the accuracy of its prediction. The proprietary modeling software is tested with actual energy use data from the building. The methodology is adapted from the methodology documented in the Model Analysis Report, which may be used to evaluate a model's performance for multiple use cases including performance tracking, anomaly detection, and demand response. However this protocol is designed to test a model's energy use prediction capability for the candidate building, as it is a key element of estimating savings.

Results of this test are pass or fail based on accuracy requirements developed for a particular example. This protocol is written to provide flexibility in its application, by describing alternatives and options for the test procedures whenever possible.

SCOPE

The scope of this testing protocol includes any software that uses measured whole-building energy use and other data such as corresponding weather data, building characteristics, or operational schedules as deemed relevant by test administrators to predict whole-building energy use, and ultimately to estimate savings. This protocol tests the accuracy of the baseline model predictions without requiring review of modeling algorithms or source code, thereby eliminating the need for in-depth technical assessment of model algorithms while protecting a software vendor's proprietary algorithms and programming.

PROCEDURE

Under this testing procedure, an individual building data set of energy use and independent variables is sent by a prospective utility program sponsor, its third party agent, or a customer ("test administrator") to a participating software vendor, along with instructions for completing the test. The vendor develops energy use predictions for the required time interval and duration and submits its estimates back to the test administrator. The test administrator evaluates the vendor's energy use predictions for accuracy relative to a specific project

application. The accuracy requirements must be developed by the stakeholders, guidance for developing these requirements are beyond the scope of this protocol.

Depending on test sponsor requirements, identifying information about participating customers such as building address, owner, building size, space use type, or other feature, may not be provided as part of the field test in order to assure test integrity. In addition, the energy data may be transformed through multiplication by an undisclosed factor.

It is recommended that a vendor's modeling software be tested with the Model Prequalifying Protocol prior to implementing the Field Test Protocol; however this is not absolutely necessary to the outcome of the field test.

The field test procedure is outlined in Figure 11. Each step in the procedure is described below. For some steps, alternative procedures and options that provide flexibility in adapting the field test to their circumstances are described.

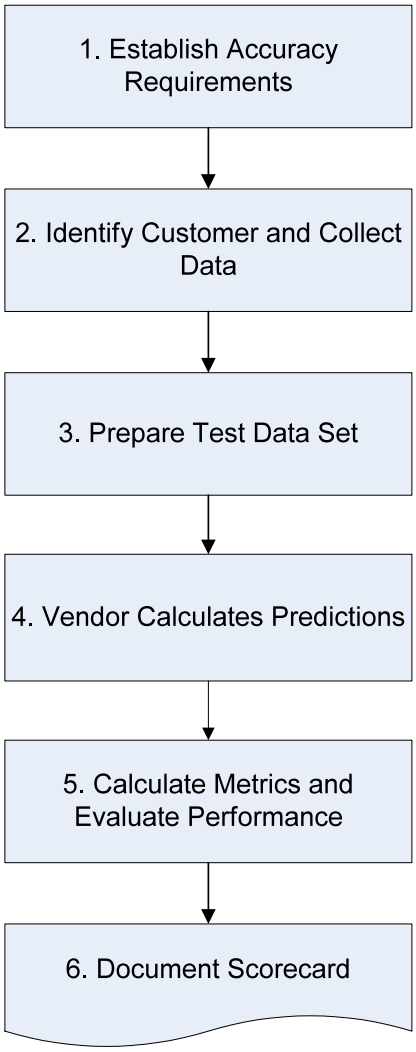


FIGURE 11. FIELD TEST PROCEDURE.

1. ESTABLISH ACCURACY REQUIREMENTS

For the whole-building savings estimation use case, the uncertainty in the savings estimation is directly proportional to the uncertainty in the baseline model's energy use prediction. Model predictions have errors, and after energy efficiency improvements have been made to a building, there is no way to know what the model error is in predicting baseline energy use, because there is no longer a way to measure the relevant baseline. "Uncertainty" is the general term to use when we are predicting a value we cannot measure. The predictions will differ from reality but we do not know exactly how much. From the Model Analysis Study we have some knowledge of the statistical distribution of the errors in that population of buildings. The Model Prequalification Protocol describes how to quantify the uncertainty for each software model for other populations of buildings.

ASHRAE Guideline 14 adopts the view that one uses the term *error* when the exact value is known, while *uncertainty* is used when no such knowledge is available.[1] The field test is set up so that the energy use in the prediction period is known, therefore *error* is the relevant term for evaluation under this test. Note that the parameters of this test are limited to the test conditions, which include building energy use and weather during the training and prediction periods. Building operations may change for other time periods, and this may cause different results for the model on the same building.

Specific project requirements for prediction accuracy may be developed from the energy efficiency project intended for the building. For illustrative purposes only, consider a typical retro-commissioning project in a medium-sized commercial office building. For such a project, 15% of the building's annual electric energy use is expected to be saved.[2] In this case, project sponsors have determined that the maximum allowable error in the baseline energy use shall be no more than half of the estimated savings, or 7.5%. This criterion will be used to assess the model's predictive accuracy for this building.

2. IDENTIFY CUSTOMER AND COLLECT DATA

The field test requires at least two years of energy use and independent variable data for each individual building. The time interval of measurement may be as frequent as 15 minutes, as is common with electric energy data from time-of-use and smart meters, hourly, or daily

[1] American Society of Heating, Refrigeration, and Air-conditioning Engineers (ASHRAE) Guideline 14-2002: Measurement of Energy and Demand Savings, Section B2.2, p. 102.

[2] Mills, E., H. Friedman, T. Powell, N. Bourassa, D. Claridge, T. Haasl, and M.A. Piette. 2004. "The Cost-Effectiveness of Commercial-Buildings Commissioning: A Meta-Analysis of Energy and Non-Energy Impacts in Existing Buildings and New Construction in the United States." LBNL-56637

intervals, as is common with natural gas smart meter data. Two years of data are required in order to set up training and prediction period scenarios, as described in the next step.

Independent variable data is required for software vendors to develop their energy models. This data includes parameters that have significant influence on energy use in a building. Such parameters include: ambient weather, solar load, operation schedule, and building occupancy. Ambient weather, including dry-bulb temperature, relative humidity, dew-point, barometric pressure and other weather variables, are typically the most accessible source of independent variable data. Reliable sources of solar data are less common. For one building, operation schedules should be readily obtainable. Actual building occupancy may be available from some buildings.

Test administrators should collect the independent variable data, however in the case of ambient weather; they may elect to provide only the building zip code or climate zone so that software vendors may collect it.

The test sponsor may consider other building information such as building size, number of occupants, or space use type to be included, however if customer identity must remain confidential, test sponsors must be careful not to divulge too much information that allows the customer to be identified.

3. Prepare Test Data Set

The participating vendor's energy use prediction software requires a set of data to train its models and independent variable data for predicting energy use. These predictions are compared to the actual energy use through the use of the evaluation metrics described in Step 4 below. This test is primarily concerned with the accuracy of energy use predictions for a subsequent year, therefore prediction periods of 12 months are required. Alternatively, model predictive accuracy for quarterly or monthly periods may be included in the test. The notation T:P is used to describe training and prediction period combinations, where T is the training period duration and P is the prediction period duration.

Test administrators may also desire to know whether a model can accurately predict energy use for a subsequent year based on shorter training periods, such as three and six months. A school of thought holds that the most recent building energy use patterns prior to a prediction period are the most likely to be representative of the prediction period. Therefore, shorter training periods may be used in the test and compared with results from longer training periods. For each T:P scenario, the concurrent weather and other independent variable data must be provided.

Note that the number of model runs multiplies for each T:P scenario analyzed. Standardized input and output file formats and file naming conventions are imperative, so that the multiple data sets may be processed and organized in a short amount of time.

4. VENDOR CALCULATES PREDICTIONS

The field test assumes that software vendors will receive the test data, run their models, and provide their predictions to the test administrator, which is similar to Path B in the Model Prequalification Protocol. For one building at a time, it is not practical for the test administrator to operate the vendor's software according to Path A in the Model Prequalification Protocol. The key attributes of the Field Test Protocol are:

1. Vendors operate their own software and provide predictions to the test administrator. This requires that they are provided energy and independent variable data. The test sponsor will determine what independent variable data, in addition to energy data, to make available for the vendors. For example, the test sponsor may make available weather station data or vendors may be required to collect this data independently. Customer-specific information such as zip code, street addresses, city, space use type, or building size may also be provided, but test sponsor requirements with regard to customer data privacy and confidentiality may limit the types of data available. In order to protect customer identifying information, the energy data may be transformed, although caution should be taken to ensure that any transformation does not compromise the quality of model test results.
2. Vendors are not required to provide their software, only its predictions of energy use. Therefore vendor's intellectual property is protected. Vendors must identify the software version being tested.
3. This test does not control for manual intervention in the software on the part of vendors. Ideally, the models should be automatically run, with energy predictions resulting purely from the models programmed algorithms. However there are no controls to assure automatic running of the vendor's models.
4. Vendors will be provided a data set for use in training their models and only independent variable data (not energy data) for the prediction period. Vendors are free to use additional information they deem appropriate in their analysis.
5. Example input data files in ASCII format will be provided. Explicit instructions and example files for vendors to return their energy use predictions will also be provided.

Test administrators may elect to compare the software vendor's model performance with that of one or more public domain models. Brief descriptions of some common public domain models have been provided in the Model Prequalification Protocol and Model Analysis Report. As described previously, the quality of public domain model predictions varies with the sophistication of their approach. Comparing energy use predictive capability of a vendor's

software to that from a public domain model can yield insight about the 'modelability' of the specific building, and about the relative quality of the vendor's software predictions.

If the comparison with public domain model performance is desired, the test administrators must set up and run the public domain models to generate the energy use predictions to be used in the comparison.

5. Calculate Metrics and Evaluate Performance

The primary goal of the field test is to evaluate the proprietary software's ability to meet the performance requirement established for the test. While the prequalification test evaluates whether the model meets requirements for a group of buildings, the field test evaluates whether a model meets requirements for a specific building. Therefore elements of the field test should be similar to that of the prequalification test, including evaluation metrics and T:P scenarios.

For this test, the primary metric of interest is the *absolute percent bias error* (APBE) as described by Equation 1 in the Model Prequalification Protocol. The APBE will be quantified for a 12 month training period and 12 month prediction period, and compared to the specific project requirements to determine whether the software passes or fails the field test.

The model is evaluated by comparing its APBE for a 12:12 scenario with the maximum error allowable by test administrators. For the illustrative use case, this error was specified to be 7.5% of baseline energy use.

Note that a fail result does not necessarily imply poor software performance, as the software may achieve a pass result for another building.

Test administrators have several options in conducting the field test. These include additional evaluation metrics, model predictive capability with shorter training periods, and comparative performance with public domain models.

Additional evaluation metrics of interest help evaluate model performance over shorter prediction periods. For example test sponsors may require more frequent savings reporting milestones after energy efficiency measures have been implemented. Therefore the model's predictive accuracy over a three-month quarter, or each month may be of interest. Test sponsors may elect to include *quarterly MAPE* (Equation 2), or *monthly MAPE* (Equation 3).

When day-to-day tracking of energy use is important, a model's daily predictive capability is important. The evaluation metric *daily n(RMSE)* (Equation 4) may be included in the test.

The predictive capability of a model based on shorter training periods may be evaluated through test scenarios with three and six month training periods. To determine these impacts, test administrators must prepare additional data sets for vendors to run these models.

Proprietary model predictive accuracy may be compared to that of public domain models. This requires a public domain model to be selected (Section 6 of the Model Prequalification Protocol) and run with the same training and prediction period scenarios as the proprietary models, and evaluated with the same metrics.

Finally, a good quality control task to include is to chart the model's energy use predictions and the actual energy use over the prediction period. This will provide a visual confirmation of model performance to support the evaluation results.

6. Document Scorecard

The pass/fail results of the Field Test Protocol may be documented in a simple report which describes the conditions of the test, the maximum error requirement, and the results of the test. If multiple evaluation metrics and T:P scenarios are included in the field test, the results may be summarized in a table. An example table format is provided in Table 5, where example results are provided for APBE in each T:P scenario only. Figure 12 and 13 provide examples of the prediction period actual energy use and a model's energy use predictions for the Time-of-Week-and-Temperature model and the Mean Week model.

Table 5. Results of Field Test

T:P Scenario	Metric	Required Score	Model Score	Mean Week Model Score	Pass/Fail?
12:12	APBE	7.50%	6.82%	9.30%	**Pass**
	Quarterly MAPE	N/A	N/A	N/A	N/A
	Monthly MAPE	N/A	N/A	N/A	N/A
	Daily n(RMSE)	N/A	N/A	N/A	N/A
6:12	APBE	7.50%	7.37%	10.40%	**Pass**
	Quarterly MAPE	N/A	N/A	N/A	N/A
	Monthly MAPE	N/A	N/A	N/A	N/A
	Daily n(RMSE)	N/A	N/A	N/A	N/A
3:12	APBE	7.50%	8.21%	11.74	**Fail**
	Quarterly MAPE	N/A	N/A	N/A	N/A
	Monthly MAPE	N/A	N/A	N/A	N/A
	Daily n(RMSE)	N/A	N/A	N/A	N/A

Figure 12. Prediction period energy use: actual and predicted.

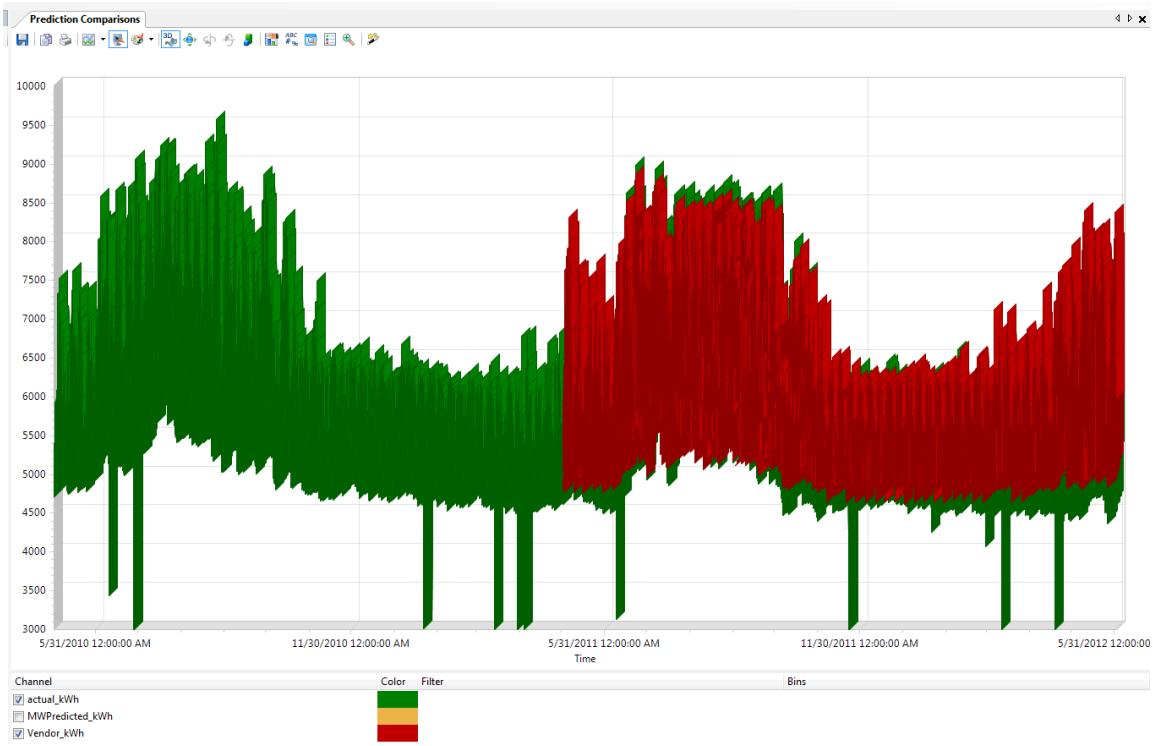
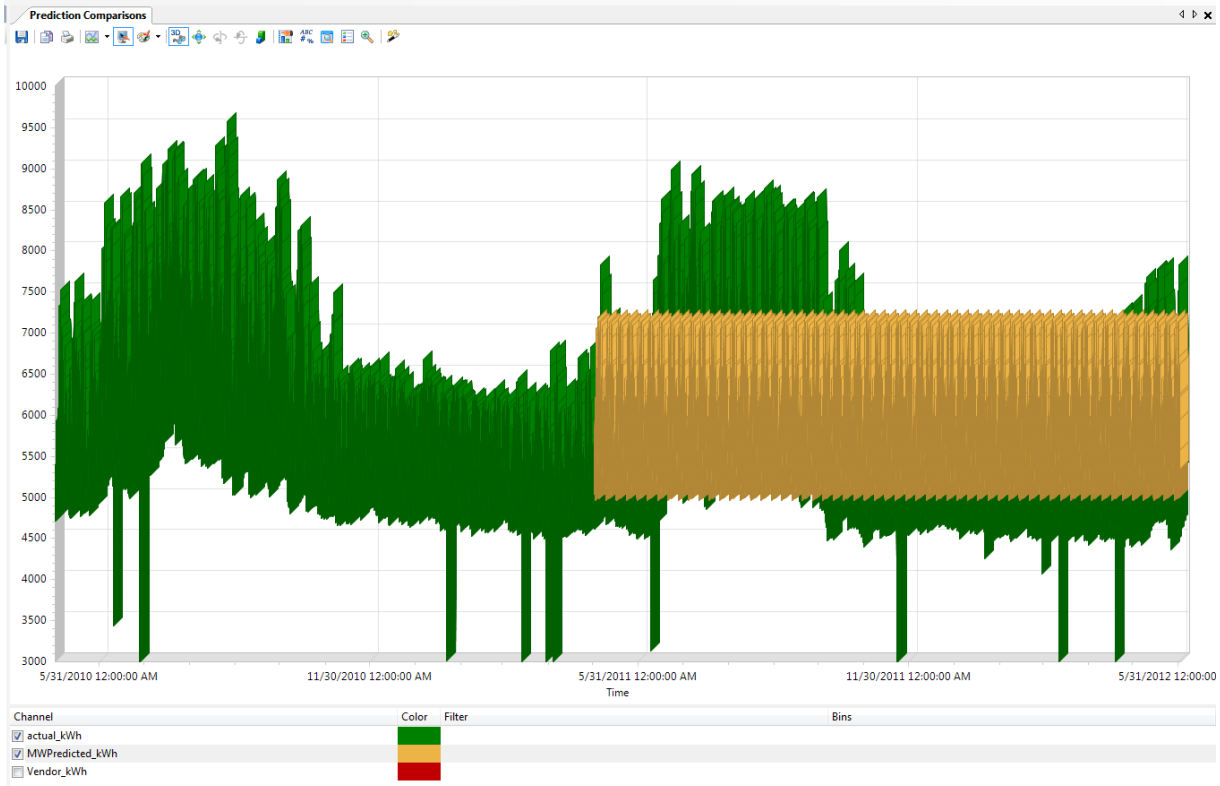


Figure 13. Prediction period energy use: Actual and Mean Week Model



REFERENCES

American Society of Heating, Refrigeration, and Air-conditioning Engineers (ASHRAE) Guideline 14-2002: Measurement of Energy and Demand Savings.

IPMVP: International Performance Measurement and Verification Protocol: Concepts and Options for Determining Energy and Water Savings, Volume 1. Efficiency Valuation Organization, January 2012

Kissock JK, Haberl JS, and Claridge DE. Development of a Toolkit for Calculating Linear, Change-point Linear, and Multiple-Linear Inverse Building Energy Analysis Models, ASHRAE, 2002

Matthieu 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(3):507–518, September 2011

Mills, E., H. Friedman, T. Powell, N. Bourassa, D. Claridge, T. Haasl, and M.A. Piette. 2004. "The Cost-Effectiveness of Commercial-Buildings Commissioning: A Meta-Analysis of Energy and Non-Energy Impacts in Existing Buildings and New Construction in the United States." LBNL-56637