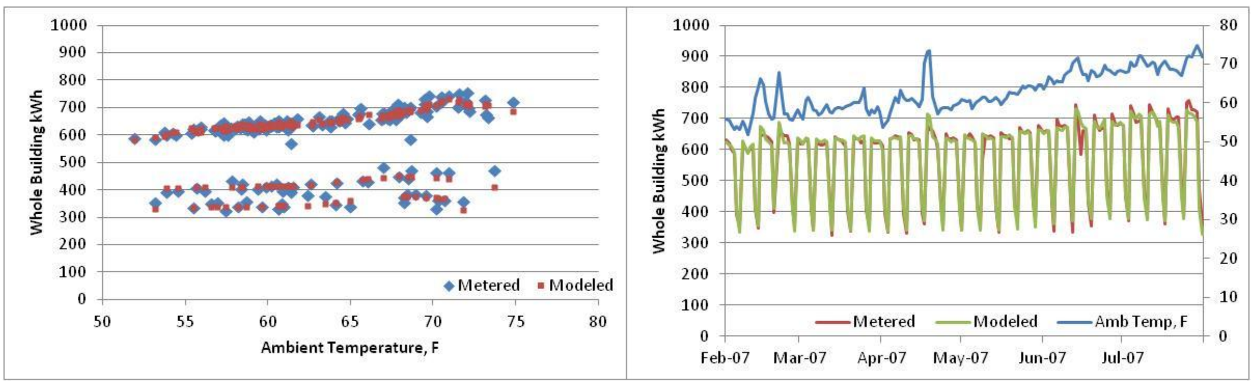
# INDEPENDENT TECHNICAL REVIEW OF THREE REPORTS RELATED TO COMMERCIAL BUILDINGS EMIS BASELINING SOFTWARE FUNCTION TESTING METRICS AND PROTOCOL

*ET Project Number: ET 12PGE5312*




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

**Prepared By: Agami Reddy**
**SAM Associates**
**10811 East Palm Ridge Drive**
**Scottsdale, AZ 85255**

**Issued: November 15, 2013**



Pacific Gas and Electric Company®

## 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 ET12PGE5312. SAM Associates conducted this assessment of the technology evaluation reports submitted by the contractors on behalf of Pacific Gas and Electric Company with overall guidance and management from Leo Carrillo. A Technical Advisory Group (TAG) was also formed which provided valuable advice and different perspectives in various facets of this research. For more information on this project, contact lmcz@pge.com.

## 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.

# About This Project

Energy Management and Information Systems (EMIS) are being increasingly installed in commercial buildings. These systems not only involve hardware capability to gather a data stream equivalent to utility-meter interval data at sub-hourly intervals, but also incorporate software solutions that allow translating this data into valuable information for building operators, owners and building services companies. For example, several vendors offer software solutions which allow baselining capability useful for assessing utility conservation program such as estimating energy savings of implemented energy conservation measures and practices with little or no need for on-site data collection. However, one barrier to realizing this opportunity is the lack of an established method and performance standard for testing the baseline modeling functionality of both open-source and proprietary EMIS software.

To overcome this technical barrier relating to the software (as against the hardware) capabilities, PG&E commissioned a research project entitled: "EMIS Software Baseline Modeling and Savings Estimation Functional Testing: Evaluation Framework, Open Source Model Performance, and Suggested Baseline Modeling Testing Protocols," (or "EMIS Baselining Project" for short). This project developed an evaluation methodology, performance metrics and test protocols by which the software capability of open source modeling tools can be evaluated.

Five technical reports were generated as outcomes of this project:

(i) Efficiency program implementation and consulting engineering firm PECI Inc. produced a "white paper" to explore the potential value to the energy efficiency industry of EMIS software savings estimation feature as well as the research questions that are relevant to quality testing of baselining functionality for estimating savings
(ii) Lawrence Berkeley National Laboratory (LBNL) with a consulting engineering firm Quantum Energy Solutions and Technologies (QuEST) documented an analysis of several open source energy modeling algorithms that identified several key performance indicators for use in model testing and evaluation.
(iii) QuEST and LBNL, with input from the project sponsor PG&E, developed a set of test protocols for use by efficiency program administrators that builds on the modeling analysis study.
(iv) Agami Reddy of SAM Associates, an engineering consulting firm specializing in building energy conservation and renewable energy applications, provided an independent technical review of the white paper, the modeling analysis and the test protocols (i.e., reports (i), (ii) and (iii)).
(v) The final report, produced by QuEST and LBNL, is scheduled to present the results of a demonstration of the software test protocol and performance metrics using a number of open source modeling algorithms and proprietary, third party software.

This report is item (iv) in the list above.

# ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| APBE | Absolute Percent Bias Error |
| ASHRAE | American Society of Heating, Refrigerating, and Air-Conditioning Engineers, Inc. |
| ECM | Energy Conservation Measure |
| EE | Energy Efficiency |
| EIS | Energy Information System |
| EMIS | Energy Management and Information System |
| ESCO | Energy Services Company |
| HVAC | Heating Ventilation and Air Conditioning |
| IPMVP | International Performance Measurement and Verification Protocol |
| LBNL | Lawrence Berkeley National Laboratory |
| MAPE | Mean Absolute Percent Error |
| M&V | Monitoring and Verification |
| MBE | Mean Bias Error |
| nRMSE | Normalized Root Mean Square Error |
| O&M | Operations and Maintenance |
| PG&E | Pacific Gas and Electric Company |
| QuEST | Quantum Energy Services and Technologies, Inc. |

# FIGURES



# TABLES

# CONTENTS

# EXECUTIVE SUMMARY

## PROJECT DESCRIPTION

Energy Management and Information Systems (EMIS) are being increasingly installed in commercial buildings. These systems not only involve hardware capability to gather a data stream equivalent to utility-meter interval data at sub-hourly intervals, but also incorporate software solutions that allow translating this data into valuable information for building operators, owners and building services companies. For example, several vendors offer software solutions which allow baselining capability useful for assessing utility conservation program such as estimating energy savings of implemented energy conservation measures and practices with little or no need for on-site data collection. However, one barrier to realizing this opportunity is the lack of an established method and performance standard for testing the baseline modeling functionality of both open-source and proprietary EMIS software.

Through its Customer Energy Efficiency Emerging Technologies Program, PG&E commissioned a research project entitled: "EMIS Software Baseline Modeling and Savings Estimation Functional Testing: Evaluation Framework, Open Source Model Performance, and Suggested Baseline Modeling Testing Protocols," (or "EMIS Baselining Project" for short), to address these and other technical barriers in an overall effort to evaluate the value of predictive analytics in determining energy use baselines and estimating energy savings from efficiency projects in commercial buildings. A primary goal of the project was to develop an evaluation methodology, performance metrics and test protocols, for interested utilities and their regulators to consider when formulating their own performance criteria based on their respective tolerances for model prediction uncertainty and error. A secondary goal was to demonstrate the application of the methodology using monitored data from a large number of buildings and relying on public domain statistical models. Other issues related to individual customer suitability and relative prediction error at the program portfolio level were also investigated.

Three primary reports were generated as part of this project. The objective of this document is to provide an independent technical review of these reports in terms of meeting the stated objectives, technical rigor and completeness, and to identify additional research directions necessary to achieve the higher level goals of this project.

## SUMMARY OF FINDINGS

PECI Inc. articulated the **potential value** to the energy efficiency industry of EMIS software savings estimation feature, and identified **research questions** that are relevant to quality testing of baselining functionality for estimating savings.

QuEST with input from LBNL and PG&E, developed a methodology to evaluate/test proprietary EMIS software modeling capability in terms of whole-building baseline energy prediction accuracy which could eventually lead to a standard.

The testing was subsequently broken up into two different protocols: (a) a test protocol, called '**prequalification test**", meant to pre-qualify a particular EMIS software tool so that the vendor could bid on certain types of energy efficiency (EE) evaluation programs, and (b) a separate protocol, referred to as "**field test**" meant to ascertain whether the EMIS software tool is appropriate for a particular (or a small set of) building(s) where certain energy efficiency measures are to be implemented. The overall assessment is that both protocols suggested are very promising and appropriate for what they are meant to achieve. The report does point out that the protocols are to be viewed as work in progress documents which needs to be refined and vetted by followup activities.

LBNL with input from QuEST documented the results of a comprehensive analysis involving numerous data sets of actual commercial buildings and using several open source energy modeling algorithms that identified several key performance indicators for use in model testing and evaluation (**functionality testing and model analysis**).

This study provided a solid statistical basis for the credibility of the pre-qualification evaluation protocol proposed. Many of the key issue or steps of the protocol were investigated. Bounds of performance accuracy achievable with fully automated open source baseline modeling algorithms have been established by this research, which is a major contribution to existing knowledge, and provides the necessary quantification in support of specific use case applications based on whole building baselining data.

The two EMIS baseline software evaluation protocols proposed, along with the background report with research questions, and substantiated by the statistical analysis in support of the pre-qualification test protocol, have fulfilled the basic objective of this entire research project. Several research gaps, some minor and some needing more attention, have been identified and stated in the main body of this report within the individual sections where each of these three reports are reviewed.

The general analytical methodology is very promising, but specific criteria (as to performance metrics, for example) are needed to evaluate the quality of baseline modeling software functionality. Clear guidance as to the applicability to EE programs need to be also developed prior to widespread adoption of these protocols not only by PG&E but by EE programs nationwide. A number of follow-up research and outreach activities are also suggested before EE program managers and EE program evaluators can adopt the evaluation methodology routinely and with some confidence.

# INTRODUCTION[1]

Energy Management and Information Systems (EMIS) were historically meant to provide building automation and control of various equipment in buildings while providing limited energy tracking capability such as trending, benchmarking and providing limited energy monitoring. EMIS vendors are expanding their product offerings by including advanced energy

[1] Much of the content in this section and the next are based on project materials prepared by PG&E for internal discussions with contractors and for background material for the Technical Advisory Group.

monitoring capability which allows tasks such as remote auditing, and load disaggregation, as well as result in inducing behavioral change among building owners, operators or users through information and customer engagement platforms and products. This source of high-frequency energy data is also suited for monitoring and verification (M&V) of energy conservation measures (ECM) installed, and to track energy use or improve the energy efficiency (EE) of buildings. This data stream provides what is equivalent to utility-meter interval data at sub-hourly intervals which is very appropriate for estimating energy savings of implemented ECM and practices by comparing pre- and post-installation energy use baselines determined through empirical modeling, with little or no need for on-site data collection.

Several EMIS vendors offer software products with such baselining capability. This automated, empirical baseline modeling functionality, if proven to be sufficiently accurate, could provide efficiency program sponsors and practitioners with a valuable tool for estimating "ex-post", building-level savings from retrofit, retro-commissioning and other measures, including those enabled by EMIS software or pursued as part of comprehensive "whole building" programs. This opportunity is especially relevant for small/medium commercial buildings, for which installing dedicated instrumentation or estimating whole building or operational savings by simulations and engineering calculations, are often uneconomic or imprecise.

However, in order for empirical baseline modeling functionality of EMIS software to be proven, significant industry and technical barriers must first be overcome. One major market transformation barrier to realizing this opportunity is the **lack of an established method and performance standard for testing** the baseline modeling functionality of both open-source and proprietary software.

Through its Customer Energy Efficiency Emerging Technologies Program, PG&E commissioned a research project, "EMIS Software Baseline Modeling and Savings Estimation Functional Testing: Evaluation Framework, Open Source Model Performance, and Suggested Baseline Modeling Testing Protocols," (or "EMIS Baselining Project" for short), to address these and other technical barriers in an overall effort to evaluate the value of predictive analytics in determining energy use baselines and estimating energy savings from efficiency projects in commercial buildings. The goal of the project was to develop an evaluation methodology, performance metrics and test protocols, for interested utilities and their regulators to consider when formulating their own performance criteria based on their respective tolerances for model prediction uncertainty and error. So as to enhance the quality of the deliverables, a Technical Advisory Group (TAG) was also formed to provide input/feedback in the design and development of the evaluation methodology, criteria and metrics.

Three primary reports which were generated as part of this project:

(1) PECI Inc. report (Kramer et al., 2013) articulated the **potential value** to the energy efficiency industry of EMIS software savings estimation feature, and identifies **research questions** that are relevant to quality testing of baselining functionality for estimating savings;
(2) Building on the previous report, QuEST and LBNL (Jump et al., 2013), with input from PG&E, the project sponsor, developed a set of **test protocols** for open source

EMIS modeling software which could be used by building efficiency program administrators and program evaluators;

(3) LBNL and QuEST (Price et al., 2013) documented the results of a comprehensive analysis involving numerous data sets of actual commercial buildings and using several open source energy modeling algorithms that identified several key performance indicators for use in model testing and evaluation (**functionality testing and model analysis**).

The objective of this document is to provide an independent technical review of these three reports in terms of meeting the stated objectives, technical rigor and completeness, and to identify additional research directions necessary to achieve the higher level goals of this project.

So as to set the context of the analysis that follows, a brief description of the project rationale is provided, and the evaluation protocol which was the outcome of this entire project is summarized.

# PROJECT RATIONALE

PG&E and a number of other utilities are engaged in technology assessments of EMIS products, both as stand-alone applications and as bundled applications. However, measuring the savings effects of these products can be difficult for a variety of reasons, such as:

- The products may have diverse functions, features and product use cases
- Their efficacy often depends on human factors (e.g., operator sophistication, incentives and availability)
- There is enormous diversity in commercial buildings, including with respect to existing building conditions, end uses, and energy usage patterns

PG&E's rationale in first focusing on performance baseline determination was based on the following assumptions:

- Baseline determination is intrinsic to performance monitoring, including energy savings quantification.
- Baseline model estimation error is a major source of savings estimation error. They produce or enable savings at a whole building level, but often at a level that falls short of the minimum 10% measurement and evaluation guideline under the International Performance Measurement and Verification Protocol (IPMVP, 2010).
- The technical robustness of baseline modeling software functionality can be established by evaluating the predictive accuracy of the baseline estimation models using a common test data set that includes utility-owned customer and interval meter data as well as publicly available geographic and meteorological data.
- There is no established performance standard for building energy use baseline models, including those which are bundled with EMIS products. In the absence of such standards, energy use baseline modeling can be a time- and cost-intensive exercise that limits the ability of technology evaluators to achieve sufficiently large measurement samples.

- The findings could inform: (i) the technology assessments for energy management and information system (EMIS) products and other technologies that deliver operational and/or behavioral savings; and (ii) the product specification requirements of efficiency programs that utilize energy use modeling tools to both establish energy use baselines and quantify energy savings. The target audience is California energy efficiency program administrators, regulatory authorities, EMIS technology vendors and national standards organizations.

---

# SUMMARY OF PROPOSED EMIS BASELINING SOFTWARE EVALUATION METHODOLOGY

The protocol developed could be used to evaluate the suitability of a proprietary model for use cases other than baseline energy use prediction for determining ECM savings; for example, condition monitoring, demand response, and anomaly detection. However, the focus of this project was to test a model's energy use prediction capability, and hence the entire discussion regarding protocol development is directed towards this use case. In essence, the proposed test protocol for evaluating EMIS software baselining capability entails ascertaining whether the model accuracy of a vendor's proprietary software tool meets certain model pre-specified accuracy thresholds which were previously determined from benchmark baseline models using open-source algorithms when applied to a pre-selected test set of buildings (see **Figure 1**).

The proposed evaluation methodology involves two rather distinct activities with some steps in common when applied to, say, verifying the actual savings realized from implementing a particular type of ECM in a utility EE program. Both paths involve analyzing the same whole-building energy use at sub-hourly or hourly time intervals for at least a year (if not more) against a standard set of independent variables. One activity consists of using open-source algorithms of statistical models to define acceptable model error tolerances specific to the risk criterion of the specific EE program based on the data set at hand. The other activity is meant to evaluate whether a particular proprietary EMIS software technology tool meets these acceptable criteria, and hence is pre-qualified to bid on allied types of utility programs.

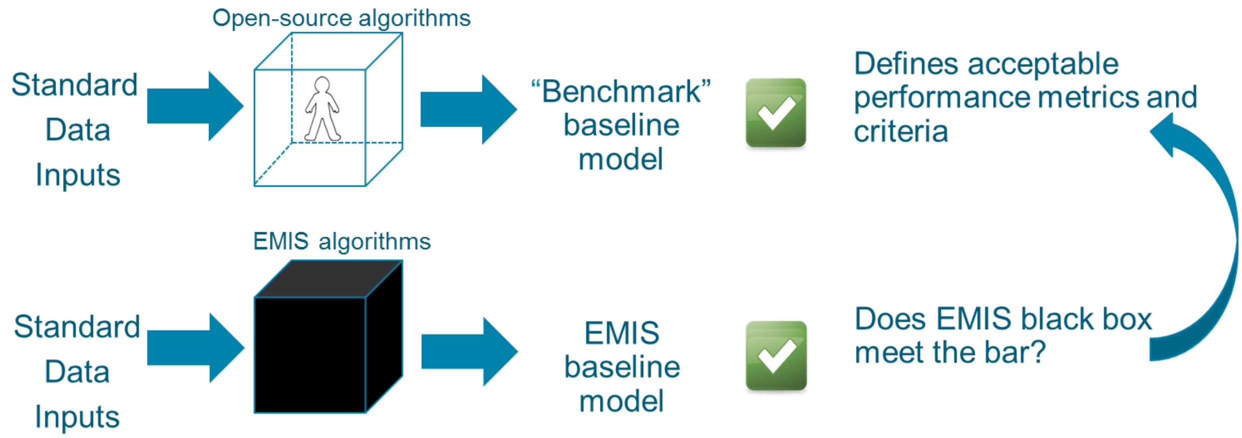



FIGURE 1. FLOWCHART OF EVALUATION PROCEDURE (FROM KRAMER ET AL., 2013)

The report by Jump et al. (2013) proposes and describes the various steps involved in the methodology to evaluate proprietary EMIS software. Actually, two protocols are proposed:

(i) A **pre-qualification methodology** whereby an EMIS software vendor who meets this test criteria specified in this protocol can become pre-qualified to bid on certain types of use case EE projects requiring the development of a baseline model from whole-building energy data (as shown in **Figure 2**). The evaluation is based on energy data from a representative sample of buildings and the use of open source modeling algorithms, and allows some flexibility in how it is administered by suggesting two options: software vendor or a third party.

(ii) **Field test qualification methodology** which applies to an individual (or a small set of) building(s) which is closely representative of the particular EE program being actually evaluated. The procedure uses blind testing similar to the protocol followed by the ASHRAE predictor shoot-out test (Kreider and Haberl, 1994; Haberl and Thamilseran, 1998), and would result in a pass/fail outcome for the specific EMIS baselining software tool being evaluated.

The protocols proposed in this report are qualified as needing more refinement and vetting, and are to be considered as work in progress. The research report by Price et al. (2013) was meant to provide a solid **statistical basis for the credibility of the qualification protocols**, to investigate several key issue or steps of the protocol, and to establish broad bounds of performance accuracy which could be achieved by fully automated open source baseline modeling algorithms.

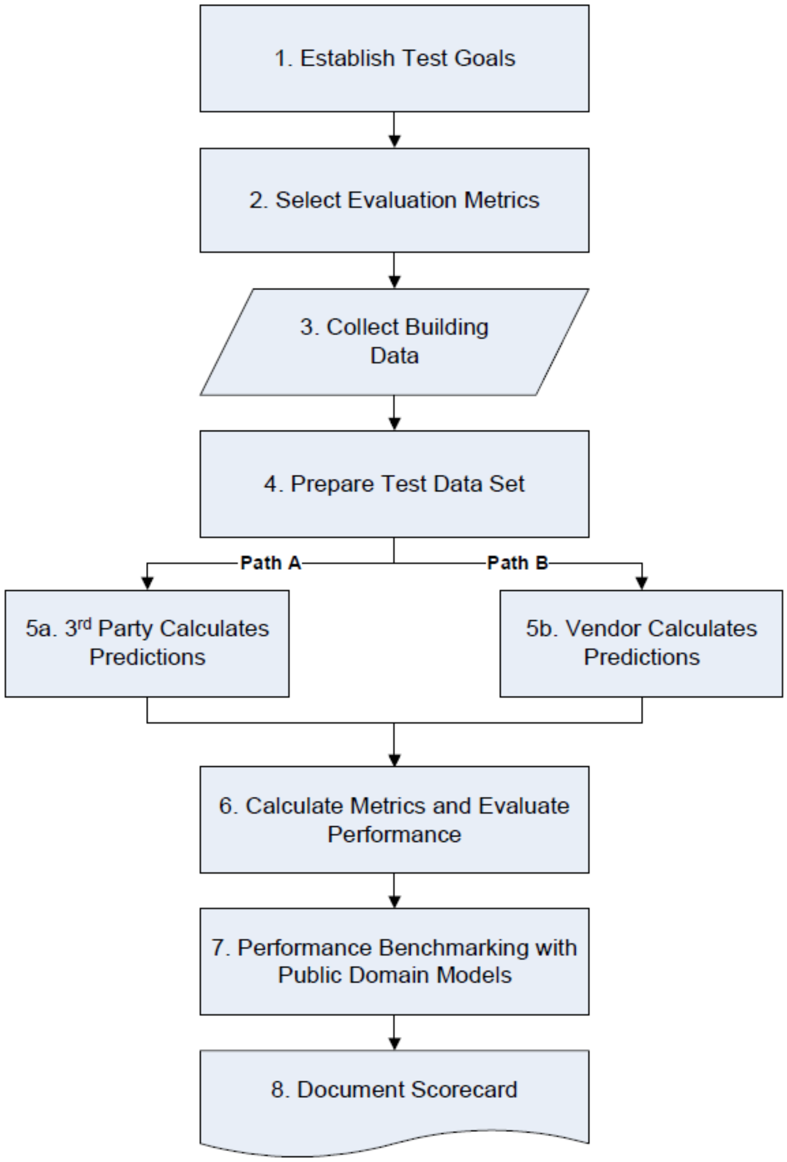



FIGURE 2. FLOW CHART SHOWING THE METHODOLOGY SUGGESTED FOR A PROPRIETARY EMIS SOFTWARE VENDOR TO PRE-QUALIFY FOR A CERTAIN TYPE OF ENERGY EFFICIENCY PROGRAM. THE EVALUATION IS BASED ON A PORTFOLIO OF BUILDINGS (FROM JUMP ET AL., 2013).

# REVIEW OF REPORT: CONSIDERATIONS FOR EVALUATING EMIS SOFTWARE BASELINING AND SAVINGS ESTIMATION FUNCTIONALITY

The high level objectives of the report by Kramer et al. (2013) were to address the **potential value** of the EMIS software technology testing protocol development project to the building EE industry, and to identify **pertinent research questions** which need to be addressed towards achieving this goal. These two objectives have been broken up into several sub-categories so as to provide more focused and specific assessment of each of them in terms of technical completeness and/or deficiencies, and to suggest ways to overcome these deficiencies, if present.

(a) General background context to EE programs based on whole building data and discuss relevant state and policy drivers:

> These aspects have been covered under the policy and market drivers section, and a pertinent overview is provided of the currently employed whole building M&V analysis approaches relevant to estimate energy savings in individual buildings. However, a literature review of how EE programs involving hundreds, if not thousands, of buildings (i.e., a portfolio of buildings) undergoing similar ECM are currently evaluated with a few typical references and description of key procedures and findings would have been useful. Identification of gaps and deficiencies of these programs would have set a better contextual stage to the current research.

(b) Importance of proper baselining using whole building interval level data and their impact on energy savings estimation:

> These issues are very well covered in the report and in Appendix A as pertinent to individual buildings. Also, **Table 1** assembling the various opportunities provides a comprehensive and excellent summary of various benefits (program cost reduction, higher accuracy, O&M benefits) as well as technical and regulatory barriers. Some discussion on portfolio of buildings is also provided, but is rather cursory.

(c) Background on EMIS technology and the emergence of their automated analytics capability:

> These aspects are covered under the sections entitled "EMIS Technology", and in Appendix B. Distinction between EIS and EMIS is adequately articulated, and a good summary description of the market trends of such EMIS systems in terms of additional analytical analysis capabilities is provided. A good discussion is also provided of how the baseline functionality of EMIS software technology can lead to various important end use functionality, most notably ECM energy savings verification and persistence. Some additional references to technical publications reporting on analysis results using the

advanced EMIS baseline modeling capability and a description of a few typical commercial EMIS products would have been useful.

(d) Articulation and justification of the need for this research project, its applications areas, and its potential impact:

The report properly identified the various stakeholders, namely utilities, ESCOs, EE program evaluators and building owners and O&M staff, and provided the necessary discussion as to their perspective. The application areas has been well covered under the section entitled "Applications for the EMIS Baselining Project", and in the section dealing with "Baselining is a stepping stone to whole building programs". However, it would have been useful to expand the potential impact treatment to include some discussion of the size of the target market sector and its energy impact as a whole. For example, the inclusion of some statistics of how many small/medium commercial buildings there are in say California, the EE programs being implemented, the dollar amount spent, the resulting energy and cost savings, and an estimate of the reduction in M&V costs in not having to install dedicated metering. Finally, a background discussion of how these EE programs are currently making use of these proprietary EMIS software analysis tools, and of the problems encountered would have been useful.

(e) Description of the software testing methodology proposed by QuEST/LBNL:

The report provided an excellent high-level perspective of the approach as illustrated by Figure 1 of this report which was drawn from Kramer et al. (2013). This is a succinct representation of the underlying evaluation methodology adopted in this research. However, the report did not elaborate on the details of the various steps proposed (such as those shown in **Figure 2** or assembled in **Table 1**) nor did it provide a proper contextual background to the various aspects of the proposed methodology. Whole building ECM savings estimation methods for individual buildings (needed for the field test evaluation protocol) are well described, while issues germane to portfolio of buildings could have been better treated.

(f) Identification of research questions pertinent to different steps of the methodology:

Several pages are dedicated to this aspect of the research in two different sections of this report. The section entitled "Key Research Questions for Baselining" deals with baselining issues only while the section on "Additional Research…" covers issues related to one specific (but important) end use application: ECM energy savings estimation. The latter section is satisfactory on the whole, and identified the important issue of how to handle improper baseline data. This erratic or large change could be due to major changes in building function or to physical changes or to improper operation.

The first section is very good on the whole and identified a large number of issues. The baselining-related section groups the research questions into four categories as shown below along with review comments:

- *baseline model inputs (further sub-divided into analysis time interval, length of baseline period, data quality, and how to select the test data set)*- The sub-divisions are logical and are quite comprehensive. The first three sections relate to individual buildings while the last one relates to portfolio of buildings. The factors identified are all valid ones, and the discussion provided is very good.

- *baseline model performance (further sub-divided into performance metrics and criteria, prediction time interval and prediction horizon)*- the discussion provided is deemed pertinent and adequate.

- *testing protocol for EMIS software tools (further sub-divided into data input specifications and protocol user)*- the treatment is also good, but some additional issues could have been identified. For example, how does one select a suite of open-source models which are deemed to provide a certain level of accuracy while not being poor in their predictive ability, thereby resulting in "loose" model tolerance thresholds? The question "can absolute thresholds be set for certification or are these specific to the EE program?" should also have been discussed. A discussion in conformance of the different routes to implement the pre-qualification protocol as described in Jump et al. (2013) would have been useful. Also, some of the difficulties in developing a test standard (say, the fragmentation and lack of technical rigor in many of the current M&V practices) should have been better articulated.

- *utility portfolio-level analysis*- the issues raised are pertinent such as scalability, how does aggregating individual building model results at the portfolio level affect model performance, and what portfolio metrics should be used for selection. However, the discussion is rather cursory. Further, several salient issues are not identified: how is risk tolerance quantitatively related to baseline model prediction accuracy levels, the process of selecting buildings to serve as a representative portfolio for pre-qualification, how the selection of the open-source modeling methods can favor or penalize certain proprietary EMIS software tools,…

(g) Identification and discussion of follow-on activities to this research:

Several follow-up activities have been identified and discussed. One of the important **research** question raised in the white paper is standard/code related: "how/can this project lead to a well-accepted standard." This question is addressed in several sections of this report, with an analogy drawn to ASHRAE Guideline 14 (2002) which proposes a prescriptive path and a performance-based path towards compliance. Further, the report has properly identified and discussed a number of issues such as persistence of EMIS-enabled savings, the handling of data discontinues, issues related to portfolio of buildings (such as how to handle buildings whose energy use is so erratic that baseline models cannot be adequately identified), and field assessment of the EMIS software evaluation methodology in collaboration with EE program evaluators and implementers.

The overall assessment is that the PECI report is quite comprehensive in terms of individual building-related issues, and contains many of the issues relevant to the project research objectives. The overall structure of the report is good, though the sections on research questions could have been organized to follow more closely the two test methodologies meant to qualify proprietary EMIS software proposed in the Jump et al. (2013) report. The other limitation is that the report is more focused on baselining issues relevant to individual buildings while the coverage of issues relevant to portfolio of buildings is rather cursory.

# REVIEW OF REPORT: FUNCTIONAL TESTING PROTOCOLS FOR COMMERCIAL BUILDING EFFICIENCY BASELINE MODELING SOFTWARE

The primary objective of the QuEST/LBNL report (Jump et al., 2013) was to develop a methodology to evaluate/test proprietary EMIS software modeling capability in terms of whole-building baseline energy prediction accuracy which could eventually lead to a standard. The testing was subsequently broken up into two different protocols:

(a) a test protocol, called '**prequalification test**", meant to pre-qualify a particular EMIS software tool so that the vendor could bid on certain types of EE evaluation programs, and

(b) a separate protocol, referred to as "**field test**" meant to ascertain whether the EMIS software tool is appropriate for a particular (or a small set of) building(s) where certain EE measures are to be implemented.

The role of this technical review is to address the general validity and rigor of the two evaluation protocols in meeting the project objectives, to comment on the logical structure of the sequential activities/steps proposed in each protocol, and to determine whether the various sub-activities identified are well-described and pertinent issues and concepts which need further work identified.

The two evaluation protocols are based on the **predictive accuracy** over a future time period of interest of the EMIS baselining software tools rather than on the specific algorithms of the software tools or how well they have fit the data used to identify/train the models. This is a very logical and valid viewpoint. The need to maintain confidentiality of the proprietary algorithms is given adequate consideration in the formulation of the protocols given that "the software vendor will not be required to divulge or share proprietary information on their software while allowing test administrators to assess their performance" (Jump et al., 2013). The report also rightly points to the need to scrub the building data sets so as to maintain the confidentiality of the customers or identity of the specific buildings.

There are several steps and aspects in common between both test protocols, while there are several important differences as well. We will review each of the protocols separately.

## PREQUALIFICATION TEST PROTOCOL

The important steps proposed in the framework of the prequalification test protocols are shown in **Figure 2**, and have been described briefly earlier. A slightly modified version of the protocol has been generated and shown in **Table 1.** The primary activities are similar except that steps 5 and 6 of the Jump et al. report are now combined into step 6. Each of the seven activities in **Table 1** have then been broken down into sub-activities to allow for a more transparent and careful review of each of the activities.

TABLE 1. OVERVIEW OF THE PROTOCOL FOR PROPRIETARY EMIS SOFTWARE VENDORS TO PREQUALIFYING ORDER TO BID FOR SPECIFIC PORTFOLIO-LEVEL ENERGY EFFICIENCY PROGRAMS (GENERATED WITH SOME MODIFICATIONS FROM THE REPORT BY JUMP ET AL., 2013)

| Step | Activity | Description of Sub-activities |
|---|---|---|
| 1 | Test sponsors to establish project goals | 1.1 Determine magnitude of program anticipated savings for the particular type of use case scenario<br>1.2 Specify the analysis time interval over which individual baseline model predictions are to be summed (say, a month or a quarter)<br>1.3 Define risk tolerance level for the program |
| 2 | Test administrators select evaluation criteria | 2.1 Identify a set of well-accepted open source statistical baseline models<br>2.2 Determine sample size of the representative test data set (building portfolio)<br>2.3 Define metrics to evaluate the model goodness-of-fit and prediction accuracy (this is impacted by steps 1.1 and 1.2)<br>2.4 Define minimum fraction of buildings which ought to meet these metrics- this can be associated with step 1.3 |
| 3 | Test administrators collect building data | 3.1 Identify a sample of buildings reflective of portfolio to serve as representative test data set<br>3.2 Acquire energy data and needed independent variables and building data for a certain number of years (say, 2 years) |
| 4 | Test administrators prepare test data set | 4.1 Screen each building of the test data set for data quality (missing data, large spikes, abnormal time trends,…)<br>4.2 Reject buildings as appropriate<br>4.3 Resample to meet preset sample size and repeat steps 4.1 and 4.2 above |
| 5 | Test administrators generate error distributions and set thresholds using open-source models | 5.1 Select a small set of open source baseline modeling algorithms<br>5.2 Fit these to each of the buildings in the test data set and generate distributions of the error metrics of the portfolio<br>5.3 Determine baseline model error thresholds based on step 1.3 and step 5.2 |
| 6 | Generate portfolio error distributions | 6.1 Repeat step 5.2 using proprietary EMIS baselining software products; either third-party (path A) or the software vendor (path B) administers the test |

| | | |
|---|---|---|
| 7 | Test administrators evaluate results and make final determination | 7.1 Summarize performance using scorecard- compare results of open-source vs. proprietary error distributions from steps 5.2 and 6.1 using error thresholds determined in step 5.3<br>7.2 If successful, proprietary software vendor is preapproved to bid on the specific type of use case utility programs |

(1) Test sponsors to establish project goals:

This step is well described with relevant issue identified such as analysis time interval over which individual model predictions would be summed, risk level, and uncertainty of baseline model at say hourly levels. It is felt that sub-activity 1.1 of **Table 1** should have been added. The tolerable risk level expressed in terms of model uncertainty should be based on the anticipated savings fraction, and explicit recognition should have been given to this aspect as described in Reddy and Claridge (2000). Finally, some more discussion of how uncertainty would decrease as the portfolio size increases should have been included.

(2) Test administrators select evaluation criteria:

The Jump et al. report (2013) named the second step as "Select Evaluation Metrics". It would have been preferable to include other evaluation criteria as well in this step. For example, the selection of the proper set of public domain or open source baselining model algorithms would have a large impact on the pre-qualification process. The models suggested in the report are described in step 6 of the Jump et al. report. It would have been better to define them in this step. The size of the portfolio is another important consideration, and perhaps sub-activity 2.4 should also have been explicitly identified since it is determined from the risk tolerance of activity 1.3. Discussion of the mathematical relationships between these considerations is briefly given in step 3 of the Jump et al. report, but some additional discussion would have greatly enhanced the rigor of this step. The errors metrics proposed and defined are valid ones which quantify both model bias (APBE) and model variability (MAPE and nRMSE); these are well accepted by the professional community, and should be a good set for vetting the proposed protocol. Whether additional or alternative metrics are more meaningful should be evaluated as part of future work.

(3) Test administrators collect building data:

This step involves collecting energy data from a representative sample of buildings. The report provides a preliminary indication of the size of the sample based on assuming a normal distribution (a size of 102 buildings is suggested). This a preliminary suggestion, and needs to be investigated further at a future time. Other than rightly stating that the sample should be random, further insights are lacking. The selection of the representative set of buildings is an important issue and needs much more guidance than that provided in the report. It is suggested that 2 years of data be gathered for each building since the uncertainty of the model predictions requires at least this length of data. Typical independent variable data and other types of data which are needed for baseline modeling are also identified.

(4) Test administrators prepare test data sets:

The description of this step is rather perfunctory. It would have been advisable to include additional sub-activities as suggested in **Table 1**. Screening of the data, rejecting certain buildings, and resampling are practical issues which would arise during actual implementation of the proposed protocol. Some guidance would have been useful.

(5) Test administrators generate model error distributions and set thresholds using open-source models:

- Discussion of the model functions of some well accepted public domain models is adequately provided, though the selection of this particular set is not justified properly. Once the protocol is refined, this reviewer does not see the need to include so many models to serve as evaluative baselines. At most two models should suffice; otherwise, program resources may be expended in needless computing and analysis effort.

- The open-source baseline models are used to generate error distributions and set thresholds for prequalifying a particular EMIS software vendor. Work leading to standards development must be largely accepted by the appropriate professional community at large. The most basic research question is "how representative are models selected as open-source models to be used as benchmarks?" The models should not result in the thresholds being set too low (or too stringent either). Have these models been fully vetted (via per reviewed journal publications- internal reports and conference papers are not adequate in this regard) and accepted and used by the professional community in terms of modeling and in terms of prediction uncertainty estimation (is the cross-validation technique proposed robust enough and fully vetted?). These key issues should been flagged and discussed further in this report.

- The Jump et al. report points out that ideally "the public domain models should be run on the same set building data as the proprietary models". It also rightly stated that this may not be feasible or even advisable, and provided two reasons. This is a valid consideration, and needs more elaboration and some closure at a future date.

- The suggestion to present the model predictive error distributions in the form of cumulative error distributions is an excellent one. It provides a visually insightful comparison of open source and proprietary model algorithms, while also shedding light on the energy use behavior of the buildings in the portfolio.

(6) Third party or software vendor generate predictive error distributions of portfolio:

Two paths for administering the assessment of proprietary EMIS software vendors are proposed. Path A involves a third party contractor using precompiled code provided the EMIS software vendor to generate the predictive error distributions for the portfolio of buildings. Path B is one where these distributions are generated by the EMIS software vendor himself. Both the paths are valid ones. The issues related to both paths are well described in the report. The requirement that the EMIS software be automated without any manual intervention so as to remove analyst expertise is a very valid point when a large number of data sets are involved, and some of the possible difficulties are also pointed out. Which of these paths will ultimately be the better one to follow can only be ascertained during field implementation of these protocols in future EE programs.

(7) Test administrators evaluate results and make final determination:

The concept of summarizing the results into a scorecard format is an appealing suggestion likely to be acceptable to all stakeholders. The sample format proposed in the report seems to capture many of the evaluation results. One suggestion is that would be advisable to reduce the number of figures for better clarity and quicker comprehension by higher level program evaluators and administrators. Providing a text box which summarizes the final results and key observations would also add value to the proposed scorecard format.

## Field Test Protocol

The purpose of the field protocol is to evaluate whether a particular proprietary EMIS baselining software tool can accurately predict energy use of one (or a small set) of building(s) within the context of a large EE program evaluation. Hence, the intent is to certify or approve the software product as appropriate for baselining energy use of the **specific building.** The Jump et al. report (2013) does not provide clear reasons why such a protocol is needed, and EE programs which will require such a building-specific protocol. The few sentences of explanation provided on the need of this protocol are inadequate.

One possible explanation which comes to mind is as follows. An electric utility would like to avoid free-riders and not overpay an individual customer for ECM energy savings. On the other hand, the utility would like to avoid situations where a customer contends that he is being under-compensated for energy savings realized in his building. In such cases, the EMIS baselining software predictive accuracy becomes critical, and the field test protocol can be used to evaluate alternative EMIS baselining software tools.

The proposed field test protocol has several steps in common with the pre-qualification test protocol. The major difference is the manner in which the testing is administered. Whereas the latter involves divulging energy use of the individual buildings over the entire evaluation time period (typically 2 years long), here the testing is a one-sided blind test. The EMIS software vendor is provided with a period where both energy use data and corresponding independent variable data (this is used to identify a model), and an additional period where only independent variable data is provided. The vendors are to use the model identified with the training data set to predict energy use over the testing period, i.e., the period where measured energy use was withheld. If the model predictions fall within a pre-set accuracy level, the proprietary software vendor is approved, else he is rejected. This type of evaluation called "blind testing" has been adopted previously in the ASHRAE predictor shootout context (Kreider and Haberl, 1994; Haberl and Thamilseran, 1998) with some success. The proposed field test is deemed appropriate and valid for the intended purpose. The description provided is rather high level and lacks many of the details needed to actually implement it in practice. A more in-depth review of the two technical publications cited above as it applies to the current research project would have been useful in providing readers with some specific details.

## OVERALL ASSESSMENT

The overall assessment is that both protocols suggested are appropriate for what they are meant to achieve. The report does point out (and rightly so) that the protocols are to be viewed as work in progress documents which needs to be refined and vetted by follow-up activities. There is great value in the report, but this could have been enhanced in several ways. First, the report could have been structured better. Though the relevant aspects of the methodology are described in enough detail for the reader to appreciate the issues, the report lacks several key sections such as an executive summary, and future extensions (especially since it is "a work in progress"). The Overview section could have been broken up into several sections such as: background, problem statement, overall methodology, …

Some key high-level research issues have not be properly addresses as pointed out above. One of the important ones relates to the prequalification test: "can one define/identify a set of buildings which can serve as baselines for all utility EE programs targeted at a specific use case involving whole-building baseline modeling capability?" The use case targeted in this research is ECM savings verification. It is well know that climate variables are more influential in extreme climates rather than in mild climates (such as California). Large buildings tend to be more internal-load dominated than smaller ones. Commercial buildings can be operated with very regular or very erratic diurnal and weekly operating schedules, with several scheduling patterns in between. Given such variability, is it logical to base the pre-qualification on a set of open source models, and can one define a single set of prototype buildings deemed representative of the behavior of all commercial buildings, even if limited to one utility service territory? The Jump et al. report does raise this issue but provides no proper guidance or discussion of alternatives.

# Review of report: Commercial Building Energy Baseline Modeling Software: Performance Metrics and Method Testing with Open Source Models and Implications for Proprietary Software Testing

This technical report was meant to provide a general validation of the pre-qualification test protocol proposed by Jump et al. (2013) which has been analyzed in the previous section. It addresses, and resolves, some of the several critical research questions raised by the EMIS software evaluation protocols. Specifically, it investigated (Price et al., 2013):

1. Whether the protocol proposed is suitable to evaluate baseline model performance, both in terms of overall robustness, and relative to other models?
2. Whether the protocol is appropriate to evaluate proprietary models that are embedded in commercial EMIS tools?
3. How to pre-screen buildings to identify those that are the most model-predictable, and therefore those whose savings can be calculated with least error?
4. Whether, and how well, the public domain models perform, and what are the associated implications for whole-building M&V)?

Each of these issues has been investigated in some detail and appropriate discussion provided. The general approach adopted first involved gathering the required whole building interval data for a very large number of buildings, and evaluating different open source baseline models by generating the types of model predictive cumulative error distributions described in the pre-qualification protocol proposed by Jump et al. (2013). Though anecdotal evidence exists in numerous publications and technical reports, and several of the load variability issues in buildings over time are well known to the professional community, the approach adopted in this research is very appropriate as a means of providing the necessary insights into several key aspects of the pre-qualification evaluation protocol.

This report documents results of the statistical analysis and sheds light into building **sampling** issues needed to gather the building portfolio data. A data set of about 400 buildings were randomly identified as representative of small to medium sized commercial buildings in the PG&E service territory, and 15 min interval data from these buildings gathered over 2 years. Another data set, referred to as "volunteer" data set, was selected from lists of PG&E's customers who had previously participated in EE incentive programs. There were large differences in types of businesses, building size and energy consumption between the volunteer and representative data sets. From the volunteer data set, 55 buildings which had multi-year data were then selected for further analysis. Rigorous statistical modeling and analysis revealed that the model prediction error statistics were very similar, which is a useful conclusion in terms of the building sampling aspect.

This research also identified relevant **model predictive error metrics** (which were the ones described in the Jump et al. (2013) report) and generated cumulative distributions of these metrics when applied to the representative data sets. Such an exhaustive analysis is an excellent way of providing initial estimates of the types of model error statistics one could expect from *fully automated* baseline models, and their implication for ECM savings estimation as a use case. The careful analysis also provided insights into how these error statistics vary across five different open source algorithms.

The effect of **prescreening individual buildings** and removing those with erratic energy use which yields poor model prediction accuracy was also well investigated, and was found to be significant. Doing so would enhance reliability and greatly improve the error distribution curves upon which thresholds are set for EMIS vendor software pre-qualification. It was found that the mean error reduced from 9% to 7% with a modest screening criterion, and then down to 6.5% when a stringent screening criterion was adopted. In the last case, 90% of the buildings had errors within 10%. These numerical values are perhaps applicable to the specific representative data set selected for this research. However, insights of this sort are important to EE program evaluators and administrators who wish to adopt the underlying EMIS baselining approach to a specific end use application (which is the basic premise of the entire project initiated by PG&E). Additional research into how to perform this screening and the validity of this approach on financial implications is needed.

The issue of using **shorter length of data** period to identify baseline models was also investigated. Specifically, how model predictions deteriorated when 3 months or 6 months of data as against a whole year to identify baseline models was also been investigated to some extent. The conclusion that a 6 month training period may be as accurate as a 12 month period has important practical implications. It is, however, somewhat restrictive in its application. There have been numerous published studies in the literature regarding this aspect, and the authors should have made reference and summarized the conclusions from these papers. Generally, 6 months or even 3 months may be adequate provided they are from the proper period of the year in climates with important annual variations in weather. Much of California has mild weather, and the conclusions stated in this research may not apply for all locations in the U.S.

The issue of how **portfolio error** will be affected as the number of buildings is increased is investigated and its practical implications well discussed. Several tables of error statistics have been generated by the analysis which reveal that the relative error of the buildings in a portfolio will tend to decrease as the size increases. Though this was somewhat expected, providing actual quantitative results leads to more realistic insights germane to this research objective.

Results of an investigation into implications of baseline errors and portfolio effects for **utility incentive payments** were also reported. This was done using the bootstrap sampling method where 10 samples of 20 buildings were generated from the entire data set, baseline models fit, anticipated "savings" of 5% and 10% assumed for the entire portfolio, and the error of the estimated savings determined. Finally, the financial implications of excess payments to customers was studied. The results showed greater estimated savings than actual savings which was due to the bias in the distribution of building type used to train the baseline models. This is a very nice approach to address this general issue, but needs additional work to firm up this methodology.

In summary, the report documents several research issues which will have great value for future energy baselining and EE savings verification projects. However, some issues need further analysis. For example, how representative are the open source models selected in this analysis in general, and how valid is the methodology followed to determine model prediction error thresholds? The issue of model uncertainty prediction is based on a simplified cross-validation technique (a commonly used and well-known statistical method), and whether the manner in which it is used can provide a robust measure of model uncertainty is somewhat questionable. Further vetting of this approach and the acceptance of the building energy analysis professionals is needed. Finally, some formatting issues would have improved the report in terms of readability and ease of comprehension. For example, the executive summary is too long for the length of the main report. More use of graphs rather than tables to summarize statistical results would have been preferable.

The research report by Price et al. (2013) provides a **solid statistical basis** for the credibility of the pre-qualification evaluation protocol proposed by Jump et al. (2013). Many of the key issue or steps of the protocol were investigated. **Bounds of performance accuracy** achievable with fully automated open source baseline modeling algorithms have been established by this research, which is a major contribution to existing knowledge, and provides the necessary quantification in support of specific use case applications based on whole building baselining data.

# RECOMMENDATIONS FOR FUTURE WORK

The two EMIS baseline software evaluation protocols proposed (Jump et al., 2013), along with the background report with research questions (Kramer et al., 2013) and substantiated by the statistical analysis in support of the pre-qualification test protocol (Price et al. 2014), have fulfilled the basic research objective of this entire research project. Several research gaps, some minor and some needing more attention, were identified and discussed under the applicable sections above. The general analytical methodology is very promising, but specific criteria (as to performance metrics, for example) are needed to evaluate the quality of baseline modeling software functionality. Clear guidance as to the applicability to EE programs need to be developed as well prior to widespread adoption of these protocols not only by PG&E but by EE programs nationwide.

The following follow-on research and communication studies are proposed before EE program managers and EE program evaluators can adopt the evaluation methodology routinely and with some confidence:

- A follow-up report, to be produced by QuEST and LBNL, is scheduled to present the results of a demonstration of the software test protocol and performance metrics using a number of open source modeling algorithms and proprietary, third party software. This is a very important phase and should provide the needed vetting of the protocols as well as allow fine tuning some of the performance criteria, metrics and thresholds stipulated in the pre-evaluation protocol.

- It is not enough to upload onto a website the various reports originated from this research. More pro-active measures are needed; for example, organizing several special workshops where selected representatives from various stakeholder groups are invited, preparing technical papers and presenting them in technical conferences.
- A pathway to develop an industry standard for evaluating baseline modeling functionality of proprietary EMIS software producers should be identified.
- The proposed field protocol needs to be elaborated further, and then tested under realistic scenarios with a specific use case application in mind before evaluators can use it to make informed assessments.
- Practical assessments in the framework of actual EE programs should be undertaken to improve the various steps for the proposed protocols to be acceptable to all stakeholders. Some examples are: investigate which additional variables and information on the building would improve baseline model accuracy and whether this is practical, provide a more rigorous statistical foundation to the current method of determining model prediction uncertainty, how does one determine which set of open source models will provide a realistic baseline, how does one correct for building operation over time, …
- Issues related to assembling a portfolio of buildings also require further investigation. For example, should buildings be pre-screened for inclusion in the portfolio, and if so, what should be the criteria. Other issues relate to the size and the manner in which a suitable portfolio of buildings is assembled.
- Another important aspect which needs further study is the framing of model uncertainty limits based on estimated energy savings and financial risk.
- Finally, future studies should be initiated to extend the scope of the current research to investigating more sophisticated EMIS software functionality such as load disaggregation, EE measure identification, assessing demand reductions from demand response programs, and continuous fault detection capability.

# REFERENCES

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

Haberl, J. S., Thamilseran, S., (1998).The Great Energy Predictor Shootout II Measuring Retrofit Savings, ASHRAE Journal, Vol.; No. 1, pp. 49-56. 1998. Also Energy systems Laboratory Report ESL-PA-98-01-01, Texas A&M University, College Station, TX 1998.

IPMVP, (2010). International Performance Measurement and Verification Protocol, Volume 1. Efficiency Valuation Organization, U.S. Department of Energy, Washington DC

Jump, D., Price, P., Granderson, J., Sohn, M., (2013). Functional Testing Protocols for Commercial Building Efficiency Baseline Modeling Software, Report submitted to Pacific Gas and Electric, Project Number: ET 12PGE5312 San Francisco, September. http://www.etcc-ca.com/reports/commercial-building-energy-baseline-modeling-software-performance-metrics-and-method-testing

Kramer H., Effinger, J. and Crowe, E., (2013) Energy Management and Information System Software Technology Assessment: Considerations for Evaluating Baselining and Savings Estimation Functionality, Report submitted to Pacific Gas and Electric, Project Number: ET 12PGE1311 San Francisco, September. http://www.etcc-ca.com/reports/energy-management-and-information-system-software-technology-assessment-considerations

Kreider, J. F.; Haberl, J. S., (1994). Predicting Hourly Building Energy Use: The Great Energy Predictor Shootout – Overview and Discussion of Results, ASHRAE Transactions, Vol. 100, Pt. 2, pp. 1104-1118. 1994. Also Energy Systems Laboratory Report ESL-PA-94-06-01, TEXAS A&M University, College Station, TX.

Price, P., Granderson, J., Sohn, M., Addy, N. and Jump., D., (2013). "Commercial Building Energy Baseline Modeling Software: Performance Metrics and Method Testing with Open Source Models and Implications for Proprietary Software Testing". Report submitted to Pacific Gas and Electric ET Project Number: ET 12PGE5312, San Francisco, September. http://www.etcc-ca.com/reports/commercial-building-energy-baseline-modeling-software-performance-metrics-and-method-testing

Reddy, T.A., Claridge, D.E. (2000), Uncertainty of "Measured" Energy Savings from Statistical Baseline Models. HVAC&R Research Journal, Vol.6, no.1, January, ASHRAE, Atlanta, GA.