Deep in Data: Empirical Data Based Software Accuracy Testing Using the Building America Field Data Repository

Preprint

J. Neymark
J. Neymark & Associates

D. Roberts
National Renewable Energy Laboratory

To be presented at the 2013 Building Simulation Conference
Chambéry, France
August 25-28, 2013

NREL is a national laboratory of the U.S. Department of Energy
Office of Energy Efficiency & Renewable Energy
Operated by the Alliance for Sustainable Energy, LLC.

Conference Paper
NREL/CP-5500-58893
June 2013

Contract No. DE-AC36-08GO28308

NOTICE

This report was prepared as an account of work sponsored by an agency of the United States government. Neither the United States government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or 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 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. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States government or any agency thereof.

This report is available at no cost from the National Renewable Energy Laboratory (NREL) at www.nrel.gov/publications.

Available electronically at http://www.osti.gov/bridge

Available for a processing fee to U.S. Department of Energy and its contractors, in paper, from:

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
phone: 865.576.8401
fax: 865.576.5728
email: mailto:reports@adonis.osti.gov

Available for sale to the public, in paper, from:

U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
phone: 800.553.6847
fax: 703.605.6900
email: orders@ntis.fedworld.gov
online ordering: http://www.ntis.gov/help/ordermethods.aspx

Cover Photos: (left to right) photo by Pat Corkery, NREL 16416, photo from SunEdison, NREL 17423, photo by Pat Corkery, NREL 16560, photo by Dennis Schroeder, NREL 17613, photo by Dean Armstrong, NREL 17436, photo by Pat Corkery, NREL 17721.

Printed on paper containing at least 50% wastepaper, including 10% post consumer waste.

DEEP IN DATA: EMPIRICAL DATA-BASED SOFTWARE ACCURACY TESTING USING THE BUILDING AMERICA FIELD DATA REPOSITORY

Joel Neymark[1], David Roberts[2]
[1]J. Neymark & Associates, Golden, Colorado, U.S.
[2]National Renewable Energy Laboratory, Golden, Colorado, U.S.

ABSTRACT

An opportunity is available for using home energy consumption and building description data to develop a standardized accuracy test for residential energy analysis tools. That is, to test the ability of uncalibrated simulations to match real utility bills. Empirical data collected from around the United States have been translated into a uniform Home Performance Extensible Markup Language format that may enable software developers to create translators to their input schemes for efficient access to the data. This may facilitate the possibility of modeling many homes expediently, and thus implementing software accuracy test cases by applying the translated data. This paper describes progress toward, and issues related to, developing a usable, standardized, empirical data-based software accuracy test suite.

INTRODUCTION

Background: Why We Are But Not Where, or, Where We Are But Not Why?

Software accuracy tests play a vital role in the continuous improvement of residential building energy analysis [Judkoff and Neymark 2006, Judkoff et al. 2010, Polly et al 2011, RESNET 2006]. Historically, established software accuracy tests are based on the Building Energy Simulation and Diagnostic Test (BESTEST) methodology [Judkoff and Neymark 2006, ASHRAE 2009]. These types of tests are included in ANSI/ASHRAE Standard 140, *Method of Test for the Evaluation of Building Energy Analysis Computer Programs* [ASHRAE 2011], and comprise idealized test suites where programs are compared to each other and/or to analytical or quasi-analytical solutions. Such deterministically oriented test cases work well for finding and diagnosing software errors; however, without direct comparisons to empirical data there is no physical truth standard of comparison with respect to overall accuracy. So, BESTEST can tell us "why we are" (or at least help diagnose why we are having errors), but cannot evaluate true accuracy relative to how a real building performs as built and as occupied.

A carefully conceived laboratory-based empirical validation study can provide both prediction accuracy testing and diagnostic capability, i.e., it addresses both the "where" and the "why." However, such procedures have been developed with only limited success. This is because such tests are an order of magnitude more expensive to develop than BESTEST-type tests, requiring substantial dedicated multi-year funding. Because of the expense of constructing facilities, such tests can be accomplished in only a limited number of climates and configurations. Also, many previously published empirical validation studies failed to empirically determine fundamental inputs (in addition to the outputs), and therefore can contain substantial bias errors [Neymark et al. 2005].

Proposed new test cases with measured audit (not laboratory) data for multiple buildings, applying a stochastic approach, provide an as-built, as-occupied energy-use target, but not much precision. Figure 1 illustrates a preliminary example of the type of accuracy observable with current data. The blue solid line and the blue dashed lines represent perfect agreement and ±40% disagreement between predicted and measured data, respectively. Here we can discern some signal (correlation of predicted versus measured energy consumption) from the noise (data scatter related to bias and random error, e.g., occupant behavior). This type of test suite addresses the "where we are, but not why." That is, we see how well we can hit the target, but when disagreement between predictions and measured data occurs, there is only limited diagnostic capability based on statistical analysis for identifying causes of disagreements.

The remainder of the paper describes development of the new empirical data-based software accuracy test.

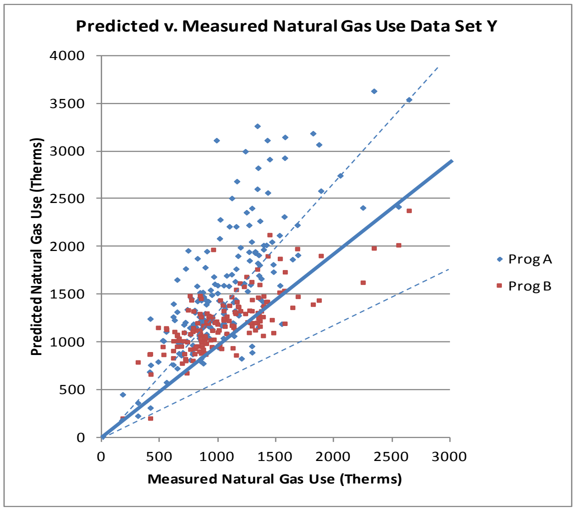


Figure 1. Predicted versus Measured Natural Gas Use from A Preliminary Study [Roberts et al. 2012]

METHODOLOGY

An opportunity is available for using home utility billing and building description data to develop a standardized accuracy test for residential energy analysis tools. Empirical data collected from around the United States have been translated into a uniform Home Performance Extensible Markup Language (HPXML) format that may enable software developers to create translators to their input schemes for efficient access to the data. It may facilitate the possibility of modeling many homes expediently, by implementing software accuracy test cases applying the translated data.

General Description

This section summarizes the empirical data-based test method fundamentals. Our objective is to assess the feasibility of developing a software accuracy test that applies currently available empirical data, for testing software as it is typically used in the field.

We have identified two possible types of datasets that could provide the basis for a test suite: population data from home energy audits, and laboratory data. Advantages of population data (audits and utility bills) are that they provide a test of programs/data as used in the field; empirical verification versus utility bills; real occupant behavior; data in many climates; ease of data use; and a relatively inexpensive test process. NREL has consistently formatted data in hand and translators for the formatted data. Disadvantages of population data are that utility dataset completeness and building description datasets vary, real occupant behavior adds uncertainty, and deterministic error diagnostics are not possible (statistical analysis is required).

Advantages of laboratory data are that physics are tested within experimental uncertainty; well-defined output data with submetering and well-defined building descriptions are provided; and direct error diagnostics are possible with precise experimental parametric variations. The primary disadvantage of laboratory data is that experiments are expensive to set up and run, so only a limited number of datasets in a limited number of climates can be collected. Also, laboratory results do not address software accuracy as used in the field, e.g., occupant behavior is automated (not real).

Because we want to evaluate software as it is typically applied, we are initially focusing on the feasibility of developing a test suite that applies current audit-based empirical data. Ideally, we would apply data where both pre- and post-retrofit building descriptions and energy use data are available for each building in a set of buildings. Unfortunately, satisfactory pre-/post-retrofit data are not generally publicly available, so we are starting with only one configuration of each building. Using the best available data, we will compare uncalibrated model results to measured utility data, where the utility data provide an empirical "truth" standard.

Ultimately, the intent is to develop an automated test that applies data for hundreds or thousands of houses, including a variety of climates and construction types. The process for achieving a full set of houses is to (1) specify one house for defining all modeling requirements; and (2) incrementally step toward a larger dataset. A number of industry participants indicated a preference for this type of approach.

Sources of Uncertainty

Sources of prediction uncertainty associated with home energy audits include systematic (bias) errors and random errors associated with (but not necessarily limited to) the following:

- Occupant behavior: i.e., operation of a house
- Building description data: including audit uncertainty, data translation uncertainty, etc.
- Weather data: measurement and location
- Utility bills: uncertainty associated with meter readings may be greater for individual monthly records than annual totals; some monthly records are estimated for some homes
- Utility bill normalization [ASHRAE 2002]: if nearby location weather data that are not coincident with the given utility bills are applied [Cummings et al. 2010]

Given the uncertainties, the method inherently tests:

- The ability to model physics within the uncertainty of typical building audits
- The appropriateness of modeling assumptions: for occupant behavior, unspecified physical components, weather data, etc.

For a given test suite, uncertainty effects may be evaluated synthetically from simulated parametric sensitivity tests, by isolating parameters either individually or in groups, and by applying reasonable ranges of variation to such parameters. For example, the sensitivity implications of changing standard operational assumptions can be assessed. Including a mix of building and equipment types in a dataset facilitates isolation of areas of greater uncertainty with stochastic analysis techniques.

Advantages of the Approach

Advantages of the test method include:

- Comparing predicted to measured energy use can demonstrate typical prediction accuracy.
- Using a population of homes quantifies prediction uncertainty across a population, allowing stakeholders to assess investment risks.
- Statistical analyses can identify model inputs that correlate with prediction errors [Roberts et al. 2012].

Limitations of the Approach

Limitations of the test method include:

- The datasets may not be representative of the broader population of homes and auditors (who collect the data). Because the data were not collected as part of designed experiments, statistical sampling procedures were not applied, so we cannot assume that the data are statistically scalable to a broader population.

- Historical data were collected for a particular purpose, using a specific data collection instrument (e.g., specific energy audit software). Auditors tend to view a house through their data collection instruments. Uncertainty may be introduced when the data are transformed to meet other needs.
- The data collected are generally limited to asset features of the home; however, often limited or no data are collected about atypical features (e.g., swimming pools). Also, only limited occupant behavior data are collected [Roberts et al. 2012].

Diagnosing Disagreements

Ideally, test cases are set up to maximize diagnostic capability – e.g., by varying individual parameters for an otherwise constant base building, as in the building physics tests of BESTEST-EX [Judkoff et al. 2010]. However, with typical house data, we cannot specify in advance each building description, and we cannot obtain measured, submetered, appliance-specific end-use data. This leads to diagnostic limitations related to:

- Raw utility data that are not submetered
- Uncertainty in the test specification (from audits, utility data, and data translations)
- Noisy parametric sensitivities – multiple parameters vary between any two houses.

Users may also imperfectly apply the test specification.

Therefore, a number of issues could cause disagreement between predicted and measured energy use, and agreement with utility bills could be attributable to compensating errors.

Limited diagnostics are possible if a sufficient number of houses are in the test suite, but require stochastic analysis methods. Such methods may correlate output to specific input variations or classes of input variations among a variety of homes, or within a variety of categories of homes, etc. Such diagnostics, rather than being deterministic, will also have an associated evaluation uncertainty. For example, if statistical evaluation of differences between predicted and measured energy use shows that heavily ground-coupled buildings tend to produce larger average errors, it could indicate a potential issue with ground modeling in the software.

Weather normalization may include some estimated utility end use disaggregation, which may be applicable for diagnostics. For example, normalizing natural gas use data requires statistically dividing the use into weather-driven and nonweather-driven portions, which provides estimates of gas use associated with space heating and gas use associated with water heating and cooking (albeit, water heating is somewhat weather dependent).

Given the availability of utility data, having different types of prediction tests may allow for enhanced diagnostics, e.g.:

- Predict both gas and electricity use
- Provide electricity use, predict gas use only
- Provide gas use, predict electricity use only
- Compare specific end-use predictions to disaggregated end-use estimations from weather-normalized utility data.

Assessing Results

A summary of potential metrics for assessing tested program results is provided later.

Existing Data and Tools

The following describes data and translation tools that can be applied for the test method.

Building America Field Data Repository (BAFDR)

The BAFDR (2013) is a collection of residential building description and energy consumption data gathered from a variety of state energy office, energy utility, and federal government efficiency and weatherization programs. The data are compiled and maintained by the National Renewable Energy Laboratory (NREL). As of this writing, the BAFDR includes about 1400 homes. The data include monthly electricity and natural gas billing records (personal data and specific addresses are deleted for privacy; zip codes and city locations are provided) and Home Energy Rating System (HERS) audit-based building description data [RESNET 2006] originally provided in a consistent REM/Rate software input format [REM/Rate 2013]. The data do not yet include pre-/post-retrofit data.

BAFDR data are translated to the HPXML home performance industry standard data transfer format [BPI 2013], and stored in that format. About 200 building description data records are provided for each house, along with monthly utility billing records. HPXML was chosen because industry software providers are rapidly adopting and using it to facilitate the transfer of data between their audit tools and their business systems.

The translation of external datasets from REM format into HPXML includes mapping building description data and utility billing records. A disadvantage of using the standard format is that the translation process adds some uncertainty to the building description if the translated format does not convey all the original data, or if the translation contains errors. This implies that careful attention to translator development is essential for developing the BAFDR's HPXML building descriptions, and when software developers map HPXML to their inputs.

Other Translation and Results Analysis Tools

Data translation (e.g., REM/Rate to selected software input files), weather normalization, and data analysis tools are described in Roberts et al. [2012].

Test Development Process

Initial Comparison Work

The test method builds off Roberts et al. [2012], where three residential energy modeling software tools were compared with an aggregation of several empirical datasets in a preliminary version of the BAFDR. The comparison required establishing preliminary criteria for eliminating ("down-selecting") specific houses from the BAFDR data to accommodate capabilities common to all three software tools. Figure 2 shows a generalized depiction of this process. As the data were mapped to each software tool, buildings with features that could not be modeled in a tool were dropped from the analysis. Furthermore, utility billing data that could not be satisfactorily normalized for differences in weather between the billing period and the simulation tool weather were eliminated. This process, as applied in Roberts et al. [2012], resulted in down-selection of houses for:

- Missing or anomalous utility billing data: homes are included only where satisfactory billing records are available for each present fuel type.
- Presence of features that could not be modeled by the tested analysis tools.

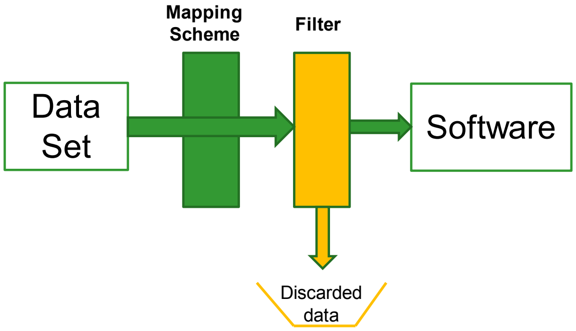


Figure 2. Test development process: data mapping and filtration

Filtered data also provide a better truth standard if weaker utility data are excluded, and less input uncertainty if weaker audit descriptions and uncommon scenarios are excluded. Less overall noise implies better diagnostic potential. The primary disadvantages of filtered data are that substantial effort is required to cross-compare data, and we risk losing variety.

To help identify potential issues driving differences between predicted and measured energy uses, multiple linear regression (MLR) analyses were employed to develop empirical models using energy use differences as the dependent variable. MLR models give indications of the inputs that most likely correlate with the dependent variable, but they do not provide absolute certainty. Key conclusions from the initial study were:

- Even if all other inaccuracies could be eliminated in an asset analysis, differences between software predictions and measured energy use would occur because occupant behavior varies from standard assumptions.
- Although occupant behavior variability is a major source of inaccuracy, it does not explain all the differences observed in the BAFDR comparisons. The remaining sources of uncertainty could be targeted to improve a given tool.

Roberts et al. [2012] track specific software improvements related to MLR diagnostics. This clearly demonstrates the potential for improving software with the test method.

Establishing a Standard Method of Test (SMOT)

Roberts et al. [2012] generated industry interest in developing an external standardized test suite that any software developer can use. Consequently, we seek to develop a SMOT based on the preliminary assessment work, with improvements primarily to:

- Provide a formal stand-alone test specification that is usable for many modeling tools, and does not require technical support from the authors of the test specification
- Provide a set of homes with the best available utility data and building descriptions.

The SMOT is being developed in collaboration with an industry working group comprised primarily of residential software developers, to establish consensus regarding the content, data format, test cases, assessment metrics, etc.

The first step for developing the test specification is to select appropriate empirical data.

The next step is to check the translation of the data into the HPXML format. This can be accomplished by comparing the resulting HPXML building description file for a given set of houses to the original database building descriptions.

After the HPXML files are vetted, the ability to use them in an automated input process is checked. This is accomplished by selecting a simulation tool(s) and developing a translator to transform HPXML to the input scheme for a given tool. Selected tool input files are then checked versus raw HPXML files, and tool input interpretations and assumptions (where directly translatable input may not be available) are also checked. These checks can be done internally by the test specification authors, and in parallel externally by the industry simulation trial participants, applying a variety of modeling tools.

After translators are developed, an iterative process of test specification development is applied. Such a process includes:

- Distributing a specified set of houses
- Obtaining and analyzing internal model and industry model results and obtaining feedback on the test spec, including identification of participant modeling assumptions
- Improving the test spec as needed

- Allowing modelers to correct translator or modeling errors, and to document corrections.

The process is repeated for additional subsets of homes until a satisfactory test suite is achieved.

As this is a new paradigm for standardized testing of building energy simulation software, we will assess its overall feasibility at key junctures. For example, the test suite may not be useful if it is too difficult for simulation trial participants to develop translators for transferring a large number of data records to their software programs' input schemes.

Getting Started: Selecting an Initial Data Subset

Criteria for initial data selection include:

- Satisfactory energy use data: This is the empirical truth standard. There must be sufficient data to ensure a full 12-month calendar period is covered (though not necessarily Jan 1 – Dec 31), and if weather-normalized utility data are applied, the data must be shown to have been successfully normalized.
- No fuel sources for which there are no time-of-use billing records: However, we recognize that secondary heating (wood, portable electric) may have been ignored in some audits, resulting in potential bias for overpredicting "conventional" space heating energy use.
- Satisfactory building descriptions: These may have greater uncertainty than energy use records because of missing data, errors in audit data entry, translation uncertainty, etc.
- Availability of thermally inefficient homes within a dataset: The modeling industry is interested in assessing retrofit candidates. Also, higher space conditioning loads provide a better signal-to-noise ratio with respect to space conditioning energy use.
- Gas furnaces in colder climates: It is often easier to separate gas-fired space heating use from the base load of other gas-fired appliances than to separate electric space heating use from electric base load, and colder climates provide a strong signal.
- Houses with common and relatively simple construction feature types (e.g., initially avoid mixed foundation types): This facilitates initial translator development by the participants and reduces the probability that software cannot properly model houses in the dataset.

Initial dataset review led to selection of data from older homes in Wisconsin collected per Residential Energy Services Network (RESNET) HERS standards as part of a statewide characterization study [Pigg and Nevius 2000]. The data have complete monthly bill data records (for more than 2 years in many homes), detailed building descriptions (REM/Rate based), and some inefficient buildings. Noticeable gaps in these data are duct areas and duct leakage rates (because the ducts are generally inside the conditioned building enclosure in Wisconsin).

Selection of "House #1"

After selecting an initial subset of houses based on the preceding criteria, we selected one relatively typical house from that subset. This house provides the basis for obtaining feedback about the initial feasibility of developing translators for going from HPXML to tested software input file formats. The initially selected house ("House #1") is located in Madison, and includes 2-1/2 years (late 1996 through mid-1999) of gas and electric monthly utility data records, where only one monthly bill was estimated. It has a gas furnace and water heater tank; space cooling and all other appliances are electric, with no supplementary fuels reported (e.g., no wood fireplace). The 2016-ft^2 house has a full conditioned basement with one floor (1008 ft^2) above ground (simple geometry). Thermal efficiency is fairly typical, with an R-19 attic, R-11 walls, R-2 windows, AFUE 90 furnace, and SEER 6.5 space cooling. The house also has a 1530 CFM50 air leakage rate, measured with a blower door test.

Test Specification

The test specification applies the methodology described above, and consists of building description data (initially for House #1), hourly weather data, monthly utility data, and simulation tool output requirements.

Building Input Description Data

The building description data are in HPXML format, and a comprehensive list of definitions of HPXML terminology relevant to the BAFDR is included in the test spec. Unlike the BESTEST-EX physics tests [Judkoff et al. 2010], which provide detailed inputs (e.g., all wall material thermal properties) and equivalent summary inputs (e.g., overall R-value), the BAFDR contains only characteristics commonly collected during an audit. For example, not every layer in a wall assembly is included, but rather the general construction (e.g., 2×4 wood frame) and the insulation R-value(s). This leaves the balance of the wall assembly details to be assumed by the software tool developer when applying the test suite to a specific tool.

Weather Data

Whether to use typical meteorological year (TMY) hourly weather data, hourly utility bill coincident weather data, or both, remains to be decided. Utility billing data must be weather normalized when TMY weather data are used [ASHRAE 2002], because the nearby TMY location weather data are not coincident with given utility bills. Use of utility bill coincident weather data eliminates uncertainty from the utility bill normalization process. However, these data may not always be available; weather data of varying measurement uncertainty and completeness for 1991 – 2005 can be obtained from NSRDB [2007]. Use of these data may also require providing a separate set of weather data for each house in a given location, depending on the time period for the utility data provided for each house. To evaluate utility bill normalization uncertainty, we may compare simulations applying real year utility bills and utility bill coincident weather data, versus TMY weather data and weather data-normalized utility bills.

Output Requirements

Initially simulation output requirements are for annual energy consumption by fuel type. Monthly (or daily) consumption may also be compared if enough tested programs can produce that output.

Modeler reports will also be included, so that we can gather feedback from users.

RESULTS

Future Results Set

The intent of initial simulation trial results is to assess the feasibility of the test suite. Initial results will include measured utility data. Results may also include internally generated results using BEopt [2013], if appropriate translators are available. BEopt is an optimization program applying established simulation engines. Other modelling tools may be compared internally to evaluate consistency of results among houses in the test suite for a given modeling tool. Industry simulation trial results may also be included, if permission is granted.

Results uncertainty analysis can be included as the number of homes in the test suite expands. A brief analysis of uncertainty caused by utility bill normalization based on noncoincident hourly weather data may also be included.

Assessment Metrics

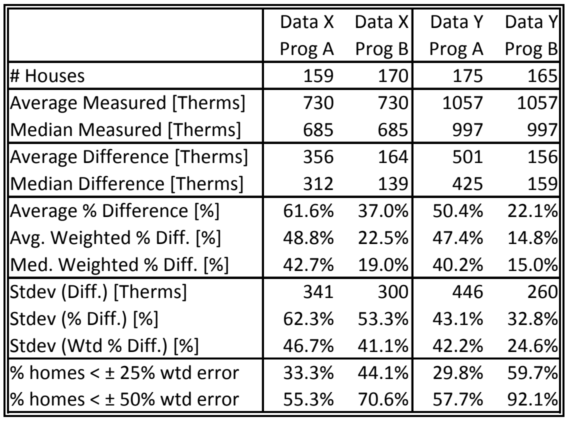
Development of assessment metrics and acceptance criteria is an iterative process, in conjunction with test specification development, reference simulation development, and industry simulation trials. Given large individual comparison uncertainty, average or median differences for a group of comparisons are important metrics, along with standard deviation and other statistical evaluation functions.

Metrics described below are an abbreviated initial set for illustrative purposes. These metrics apply preliminary data with in-situ occupant behavior, which may be considered as hypothetical data here. Example summary statistics are shown in Table 1.

The following definitions are applied in Table 1:

- Data X, Y: datasets X and Y
- Prog A, B: programs A and B
- # Houses: number of houses
- Average measured: average annual gas use
- Average difference: (Σ(pred,i – meas,i))/(# records), where pred,i and meas,i are predicted and measured use for individual record "i".
- Average % Difference: (Σ((pred,i – meas,i)/meas,i))/(# records) × 100
- Average Weighted % Diff.: (Σ((pred,i – meas,i)/meas,i x (meas,i/meas,av)))/(# records) × 100, where meas,av is the average of measured data for a given dataset; weighting addresses skewing by large differences for low use, and can also be normalized by "meas,median".
- Stdev (Diff.): Standard deviation of individual differences
- Stdev (% Diff.): Standard deviation of individual percent differences
- Stdev (Wtd % Diff.): Standard deviation of individual weighted percent differences

Table 1. Assessment Metrics: Example Comparisons



	Data X Prog A	Data X Prog B	Data Y Prog A	Data Y Prog B
# Houses	159	170	175	165
Average Measured [Therms]	730	730	1057	1057
Median Measured [Therms]	685	685	997	997
Average Difference [Therms]	356	164	501	156
Median Difference [Therms]	312	139	425	159
Average % Difference [%]	61.6%	37.0%	50.4%	22.1%
Avg. Weighted % Diff. [%]	48.8%	22.5%	47.4%	14.8%
Med. Weighted % Diff. [%]	42.7%	19.0%	40.2%	15.0%
Stdev (Diff.) [Therms]	341	300	446	260
Stdev (% Diff.) [%]	62.3%	53.3%	43.1%	32.8%
Stdev (Wtd % Diff.) [%]	46.7%	41.1%	42.2%	24.6%
% homes < ± 25% wtd error	33.3%	44.1%	29.8%	59.7%
% homes < ± 50% wtd error	55.3%	70.6%	57.7%	92.1%

measured_A_B-mmddyy.xlsx!meas_A_B-charts-X_Y(AL1188:AP203)

Table 1 provides a limited example set of metrics; other metrics that could be included are, e.g.:

- Median percent difference
- Percent of homes with < ± x% error

The data allow assessment of the relative differences (caused by bias and random errors) of the predictions when compared with each measured dataset. Average percent difference and standard deviation of the percent differences (or the weighted percent differences) allows comparison of the variation of relative differences among the datasets by normalizing for differences in measured consumption. The following is observed in the example of Table 1:

- Accuracy of program predictions
  - o Bias error (averages of predicted versus measured differences, percent differences, and weighed percent differences) is greater for Prog A than Prog B for both datasets.
  - o Random error (standard deviations of predicted versus measured differences, percent differences, and weighted percent differences) is greater for Prog A than Prog B for both datasets.
- Prediction accuracy relative to the datasets (ability of programs to model a given dataset)
  - o Bias errors (average percent difference) are greater for Data X than for Data Y.
  - o Random errors (standard deviation of percent differences) are greater for Data X than Data Y.
  - o Data Y have 45% greater annual average consumption than Data X, which implies a stronger driving function for the models, and could be contributing to the lower relative errors for the Data Y simulations.

This analysis would steer us toward choosing Data Y as the initial dataset from which to select House #1. Figures 3, 4, and 5 show ways to bin the occurrence of differences for the dataset shown in Figure 1.

Figures 3 and 4 corroborate that Program B has better agreement with measured data than Program A, with Figure 4 indicating that about 70% of Program B results, but only about 35% of Program A results, are within 30% of measured data. Figure 5 is useful because it scales differences to therms (and therefore energy cost), and shows that the largest percent errors occur for the lowest natural gas use.

The comparison plots can be produced for other datasets, to compare the ability of programs to model a given dataset (i.e., to evaluate dataset quality).

CONCLUSIONS

Accomplishments

A new methodology is being developed for evaluating the accuracy of building energy analysis computer programs by comparing simulation tool predictions to measured data, using available empirical data from the BAFDR. The method will include a test specification and assessment metrics, with the intent to standardize the method applied by Roberts et al. [2012]. Key improvements versus preliminary work are to reduce modeling uncertainty by initially selecting simpler-to-model homes that provide a strong energy consumption signal, and using HPXML as a data transfer standard. Advantages of the method include the possibilities for:

- Demonstration of typical prediction accuracy and identification of prediction uncertainty, allowing stakeholders to assess investment risks.
- Creation of a results set where statistical analyses can identify correlation between model inputs and prediction errors.

Challenges

The primary challenge is motivating software developers to write scripts for translating a large number of data records from HPXML to their models' input schemes. Although some software tool developers have already adopted HPXML for selected applications, others may find this requirement a barrier to utilizing the test suite.

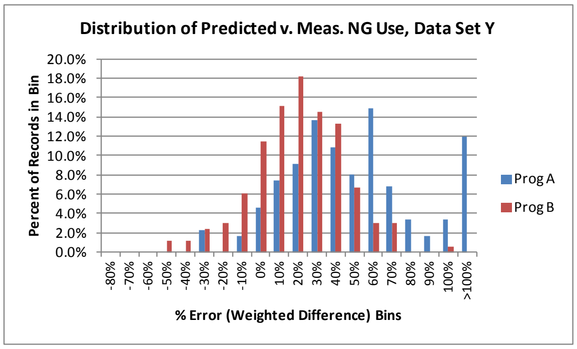


Figure 3: Binned percent differences of predicted relative to measured natural gas use

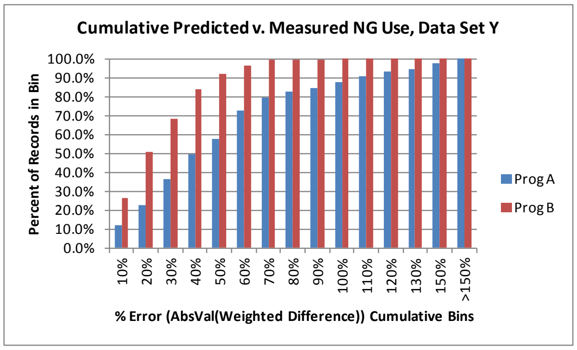


Figure 4: Cumulative binned percent differences of predicted relative to measured natural gas use

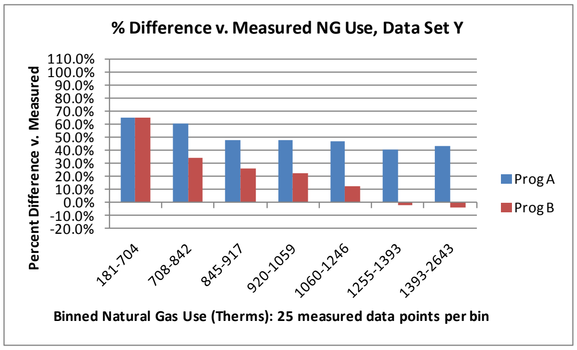


Figure 5: Average percent difference versus binned measured natural gas use

We also must be able to obtain data where predicted energy use, based on variations in building input descriptions, has a logical correlation with variations in measured energy consumption. Preliminary work provides for cautious optimism here. Roberts et al. [2012] found meaningful signal in the noise, in that they identified potential software improvements based on stochastic analysis of disagreements between predicted and measured energy use. Also, preliminary work with assessment metrics indicates some datasets exhibit better agreement between predicted and measured results than others with supporting logical causation for the better agreement.

Future Work

For the initial SMOT, we are beginning with the best available, easiest to model data. The next steps are to complete a draft test specification, run preliminary simulation trials, and obtain feedback from software developers regarding the feasibility of the test suite. If initial feasibility is promising, we plan to expand the test suite dataset to include more houses, and apply statistical and uncertainty analyses to the larger results set, as appropriate. This may also include evaluating the feasibility of using utility bill coincident weather data and evaluating utility bill normalization uncertainty by comparing results for actual utility bills and utility bill coincident versus typical published non-coincident weather data and weather-normalized utility bills.

In the longer run, we should consider how to improve testing energy savings predictions versus the current capability of BESTEST-EX. In the absence of sufficient pre-/post-retrofit empirical data, synthetic parametric sensitivity test options could involve using simulation tools to, e.g., apply parametric variations to selected BAFDR homes to tune the BESTEST-EX base case building description and reference program results to correlate better with as-built, as-occupied home energy consumption.

ACKNOWLEDGMENTS

This work is supported by the U.S. Department of Energy's Building America program.

Ron Judkoff and Ben Polly of NREL provided technical reviews and consultation throughout the project. Noel Merket of NREL and Sean Casey (formerly of NREL) provided support on the BAFDR, HPXML, and translation and analysis tools. Ren Anderson of NREL provided managerial support. Mike Kennedy of Mike D. Kennedy, Inc. provided preliminary dataset evaluation.

We appreciate the support and guidance of David Lee and Eric Werling of DOE.

NOMENCLATURE

ASHRAE: American Society of Heating, Refrigerating and Air-Conditioning Engineers.

BAFDR: Building America Field Data Repository

HPXML: Home Performance Extensible Markup Language

MLR: Multiple Linear Regression

NREL: National Renewable Energy Laboratory

RESNET: Residential Energy Services Network

SMOT: Standard Method of Test

TMY: Typical Meteorological Year

REFERENCES

ASHRAE. 2002. *ASHRAE Guideline 14-2002: Measurement of Energy and Demand Savings*. Atlanta, GA: ASHRAE.

ASHRAE Handbook Fundamentals. 2009. Atlanta, GA: ASHRAE.

ANSI/ASHRAE. 2011. ANSI/ASHRAE Standard 140-2011, *Standard Method of Test for the Evaluation of Building Energy Analysis Computer Programs*. Atlanta, GA: ASHRAE.

BAFDR. 2013. *Building America Field Data Repository.* Golden CO: NREL https://buildings.nrel.gov/bafdr/

BEopt. 2013. *Building Energy Optimization Software.* Golden, CO: NREL. http://beopt.nrel.gov/

Cummings, J.; Parker, D.; Sutherland, K. 2010. *Evaluation of Bias Issues within Regression-Based Inverse Modeling Methods Against Climate and Building Characteristics Using Synthetic Data.* FSEC-CR-1863-10. Cocoa, FL: Florida Solar Energy Center.

BPI. 2013. *Home Performance XML.* Malta, NY: Building Performance Institute. www.homeperformancexml.org/.

Judkoff, R.; Neymark, J. 2006. "Model Validation and Testing: The Methodological Foundation of ASHRAE Standard 140." *ASHRAE Transactions* 112(2):367–376. Atlanta, GA: ASHRAE.

Judkoff, R.; Polly, B.; Bianchi, M.; Neymark, J. 2010. *Building Energy Simulation Test for Existing Homes (BESTEST-EX).* NREL/TP-550-47427. Golden, CO: NREL.

Neymark, J., P. Girault, G. Guyon, R. Judkoff, R. LeBerre, J. Ojalvo, and P. Reimer. 2005. The "ETNA BESTEST" empirical validation data set. *Proceedings of Building Simulation 2005*, Montreal. International Building Performance Simulation Association.

NSRDB. 2007. *National Solar Radiation Database 1991-2005 Update:* User's Manual. NREL/TP-581-41364. Golden, CO: NREL.

Pigg, S.; Nevius, M. 2000. *Energy and Housing in Wisconsin: A Study of Single-Family Owner-Occupied Homes.* Energy Center of Wisconsin, 199-1.

Polly, B.; Kruis, N.; Roberts, D. 2011. *Assessing and Improving the Accuracy of Energy Analysis for Residential Buildings.* NREL/ TP-5500-50865. Golden, CO, USA: NREL.

REM/Rate. 2013. *REM/Rate.* (Software). Boulder, CO: Architectural Energy Corporation. www.archenergy.com/products/remrate

RESNET. 2006. *Mortgage Industry National Home Energy Rating Standards*. September 2007. Oceanside, CA: RESNET, Inc. www1.resnet.us/standards/iecc/procedures.pdf.

Roberts, D.; Merket, N.; Polly, B.; Heaney, M.; Casey, S.; Robertson, J. 2012. *Assessment of the U.S. Department of Energy's Home Energy Scoring Tool.* NREL/TP-5500-54074. Golden, CO: NREL.