A Method to Test Model Calibration Techniques

Ron Judkoff, National Renewable Energy Laboratory (NREL)
Ben Polly, National Renewable Energy Laboratory (NREL)
Joel Neymark, Joel Neymark Associates (JNA)

ABSTRACT

This paper describes a method for testing model calibration techniques. Calibration is commonly used in conjunction with building simulation software to estimate energy savings. A calibration technique is used to reconcile building simulation software predictions with building energy consumption data (typically obtained from utility bills),[1] and then the "calibrated model" is used to predict energy savings from a variety of retrofit measures and combinations thereof. Current standards and guidelines such as BPI 2400 (ANSI/BPI 2400-S-2015) and ASHRAE 14 (ASHRAE Guideline 14-2014) set criteria for "goodness of fit" and assume that if the criteria are met, then the calibration technique is acceptable. However, even with a good fit to building energy consumption data, the calibration technique may not always result in more reliable predictions of energy savings. Therefore, the basic idea here is that the building simulation software (intended for use with the calibration technique) is used to generate simulated building energy consumption data as a synthetic data surrogate for measured data against which the calibration technique can be tested. This provides three figures of merit for testing a calibration technique, a) goodness of fit to the "synthetic truth" energy consumption data, b) closure on the "true" input parameter values, and c) accuracy of the post-retrofit energy savings estimate. These three metrics provide a rigorous test by which calibration techniques can be evaluated. The paper also discusses the pros and cons of using this analytical testing approach versus trying to use real data sets from actual houses.

Introduction

Calibration is commonly used in conjunction with building energy simulation software to increase the accuracy of post retrofit savings predictions. Other terms frequently used to describe model calibration include model tuning, model true-up, and model reconciliation. Typically, residential and commercial model calibration has been implemented using monthly energy consumption data obtained from utility bills for an existing building that is about to receive an energy retrofit. Sometimes sub-metered, disaggregated, or higher frequency data is also available and used to assist the calibration process. An audit is conducted to gather information about the building needed to assemble an input file for a building energy simulation tool. A calibration technique is used to reconcile model predictions with the utility data, and then the "calibrated model" is used to predict energy savings and energy cost savings from a variety of retrofit measures and combinations thereof. Many variations on this approach exist, including some where the savings predictions are subjected to calibration instead of, or along with the model inputs.

[1] This is the most typical case however calibration can be done using any energy related measurements in the building including temperatures, and this test method is also applicable for those cases.

While it is logical to use building energy consumption data to calibrate the building simulation model, the calibration technique may not always result in a more reliable estimate of energy savings. When calibrating a large number of inputs to a limited number of outputs (mathematically this is called an underdetermined or over-parameterized problem), there can be many combinations of input parameters and values that will result in a close match to the base case or pre retrofit utility bill consumption data, so a close match is not in itself proof of a good calibration (Reddy et. al., 2006). The lower the frequency of the building performance data, or the lower the informational content of the data, the lower the probability that the calibration actually improves the model and associated energy savings predictions. Therefore, it is useful to have a method of test for calibration techniques that provides at least the following three primary figures of merit: a) the goodness of fit between the calibrated and "synthetic utility consumption data", b) how closely the calibrated input parameter values match the "true" parameter values, and c) the accuracy of the savings prediction. A useful secondary metric is "Benefit of Calibration" which compares the calibrated savings prediction to the savings prediction that would have been obtained absent calibration. Not all calibration techniques can use all of these metrics. Techniques that produce a calibrated pre-retrofit model can use all the metrics, however techniques that only calibrate the savings are limited to metric c.

Most practitioners think of validation in terms of using data from real buildings (empirical validation). A limiting factor in attempting to empirically validate calibration techniques is the lack of high quality monthly consumption data for a year before and after retrofit (higher frequency data and sub-metered data can help improve the calibration), good pre- and post-retrofit building characteristics data, local pre-and post-retrofit weather data, and the dates of the retrofit installations. Until a statistically significant amount of such data is available to researchers, an alternative analytical test method can be used in which a building energy simulation tool is used to generate its own pre and post retrofit consumption data as a synthetic surrogate for real building data. The analytical test method has some advantages over empirical tests because it facilitates diagnosis of weaknesses in calibration techniques and indicates how they can be improved. In real buildings the model inputs that represent the building contain considerable uncertainty, so it is not possible to use metric b and in real buildings it is not possible to isolate the effects of the simulation program from the effects of the calibration technique. The analytical test method facilitates "self-testing" of a calibration technique, and is useful in several ways, including: a) testing a single calibration method to see how well it works under a variety of test conditions, b) testing several calibration methods to determine under what test conditions each is best, c) investigating how much, and what kind, of informational content is needed in the synthetic data to achieve good calibrations with different calibration methods (eg. monthly vs daily vs hourly data and availability of different types of submetered or disaggregated data), d) testing with various amounts and kinds of "noise" in the synthetic data, and e) diagnostic testing. Ultimately, when data becomes available, analytical and empirical tests should be used together with the empirical tests providing bottom line validation and the analytical test method providing for diagnostics and improvement.

The analytical test method was developed for testing calibration procedures used with residential retrofit audit software however the concept could also be applied in a commercial building context. The method was initially proposed in a series of NREL reports called BESTEST-EX. BESTEST-EX described two alternative analytical test methods and a set of example test specifications that can be used in lieu of creating new test specifications (Judkoff et. al., 2010). The approach emphasized in BESTEST-EX used several reference simulation

programs to generate average synthetic energy consumption and savings data (Judkoff et. al., 2011a). Such an approach tests both the simulation program and the associated calibration technique together. The test method described in this document is different in that any given building simulation tool can be used as its own reference in conjunction with a calibration technique to test the calibration technique. BESTEST-EX introduced this concept and named it the "pure" calibration test method (Judkoff et. al., 2010, 2011a). Here, we further develop this method.

In the pure method a simulation tool generates its' own set or sets of "synthetic truth" data. Such an approach is a "pure" test of the calibration technique and does not test the physics, mathematics, and algorithms embodied in the associated simulation program. For methods to test whole building energy simulation programs see the NREL BESTEST reports at www.nrel.gov and ANSI/ASHRAE Standard 140 (Judkoff and Neymark 1995a and b, Neymark and Judkoff 2002, 2004, Neymark et. al. 2008a and b).

The "pure" method for testing calibration techniques follows the general procedures outlined below. The procedures use a number of concepts and terms explained in the "Definitions" section.

Definition of Key Concepts and Terms

Approximate input: An input that has been determined to be uncertain and sensitive. Such inputs are good candidates to test model input calibration techniques, and may also be good retrofit candidates.

Approximate Input Range: Defined range of input value uncertainty for a given approximate input.

Explicit Input: An input value selected from within a defined range of uncertainty (see approximate input range). The explicit input is the "true" input.

Explicit Model: The simulation model that contains the explicit inputs and which is used to generate the "synthetic truth" energy performance data (typically gas and electric consumption data as would appear on a utility bill for a real building).

Explicit results: Energy performance outputs from the explicit model. These outputs are synthetic data surrogates for data that could be measured in a real building (typically monthly energy consumption data).

Nominal Input: The most probable correct input, frequently obtained from an audit or by using credible sources such as the ASHRAE Handbook of Fundamentals.[2] The nominal inputs are the inputs that would have been used in the absence of calibration. In BESTEST-EX nominal inputs are sometimes referred to as physics inputs, although in rare instances they may be slightly different.

Nominal Model: The simulation model that contains the nominal inputs and is therefore the uncalibrated model. In BESTEST-EX nominal models are sometimes referred to as physics models or physics test cases.

Nominal Results: Output from the nominal model and therefore the uncalibrated results. In BESTEST-EX nominal results are sometimes referred to as physics results or physics test case results.

[2] There are use cases where nominal inputs might not be the most probable correct input such as to represent typographical or bias error in the uncalibrated input file.

Permissible data: Data types and frequencies which have been defined at the beginning of the test procedure by the test designer (test maker) as allowed to be known and used by the calibrator (test taker).

Outline and Explanation of Procedures

Figure 1 illustrates the overall conceptual approach to testing model calibration techniques and shows both the "pure" and "reference program" methods. The left column of the figure summarizes what the test designer (test maker) does and the right column summarizes what the calibrator (test taker) does. An example is given to describe the procedure. The example is for a case where the use of monthly energy consumption data to perform the calibration is assumed, however it is possible to calibrate using other kinds of measured data in buildings such as temperatures, heat fluxes, etc. Also, the example assumes that the calibration technique being tested is the type that produces a calibrated pre-retrofit model to which appropriate retrofit related inputs are then applied. While the example is typical, the test method can be used with any calibration technique that uses pre-retrofit data to improve estimation of post-retrofit savings. The description of the procedure will correspond to the boxes in figure 1.

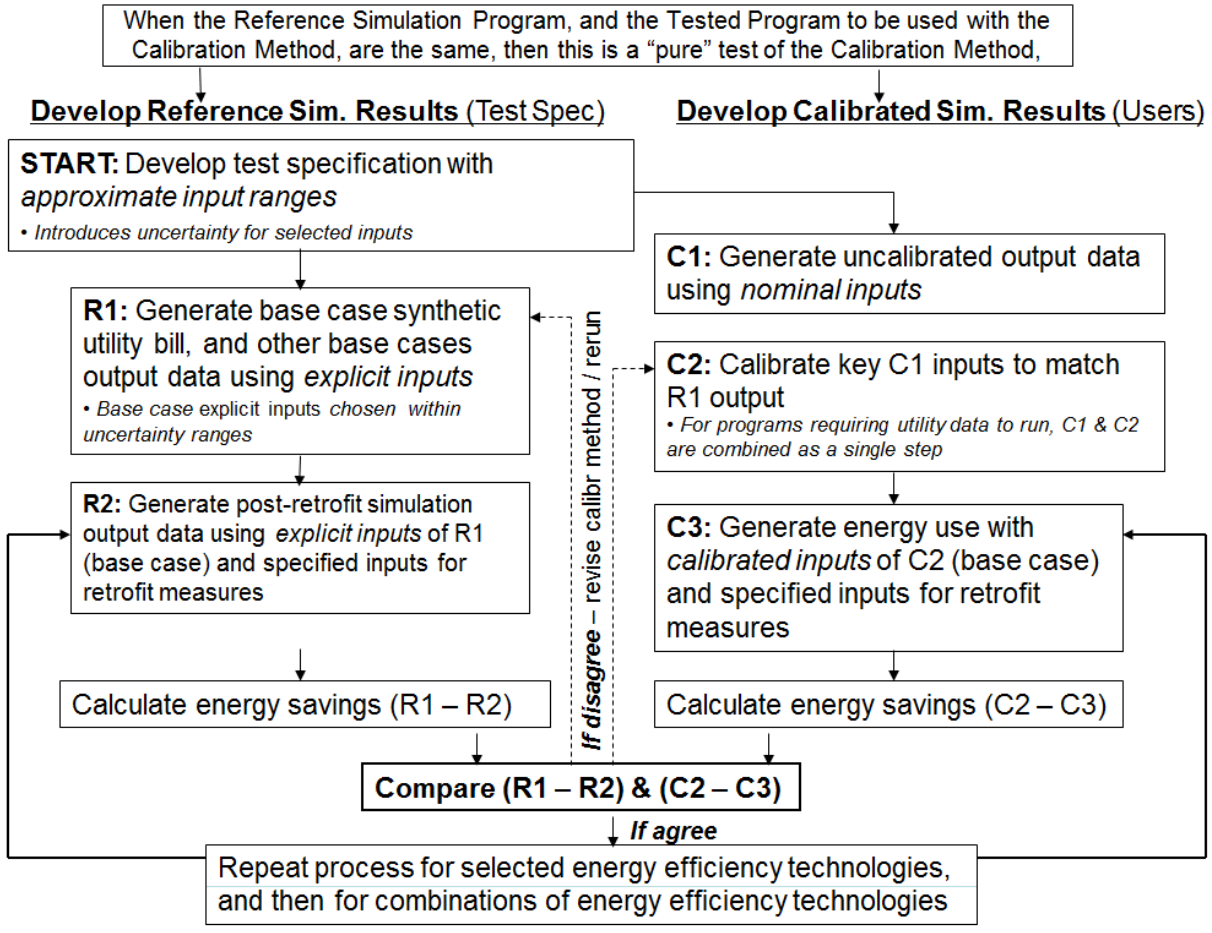

Figure 1. Conceptual Flow of Calibration Test Method. *Source*: Judkoff et al. 2011.

"Start" Box in Figure 1

The test designer defines the scope of the calibration testing. The purpose is to determine at the beginning of the test process the exact nature of the test, what information is permitted to be known and used by a human or automated calibrator, and what metrics shall be used to evaluate the test results. For example, if the scope of the calibration testing is intended to mimic a common audit and calibration scenario (e.g., the approach used in BESTEST-EX), then the only permissible data for the calibrator would be a) pre retrofit "synthetic truth" whole-building monthly gas and electric energy consumption data (no disaggregated or higher frequency data allowed for this example), b) nominal inputs representing the input data that would be collected by an auditor or modeler, and c) input uncertainty ranges for those inputs that are to be calibrated. The metrics for this example would typically be a) calibrated pre retrofit monthly and annual energy consumption data versus explicit (synthetic truth) pre retrofit monthly and annual consumption data, b) calibrated input values versus explicit (synthetic truth) input values, and c) calibrated monthly and annual energy savings versus explicit (synthetic truth) monthly and annual energy savings. If the goal of the testing is to determine the benefit of using hourly smart meter data, or sub-metered or disaggregated data, as was done in Robertson, Polly, and Collis (2013) then the test designer must define precisely what types and frequency of data are permitted to be known and used by the calibrator and what types, frequencies, and mathematical relationships of results data will be used as test metrics.

The test designer creates specifications for a pre-retrofit base case test building defining the values for nominal input parameters that a building simulation model would need. This does not have to be a real existing building, but the specification should be representative of the types of buildings for which the calibration technique will be used. If the test designer does not wish to create the test case specifications, the tests already defined in BESTEST-EX may be used.

The test designer uses the nominal model to generate nominal pre-retrofit energy consumption results. The nominal base case model can be used by the test designer to determine parameter sensitivities. This is an important step because it is most useful to test the calibration technique against parameters that are both uncertain and influential and such parameters may also be good retrofit candidates.

The test designer creates specifications for individual and packages of retrofit measures to be applied to the pre-retrofit test building. Nominal input values for the retrofit measures should be expressed as "relative" or "absolute" values depending on the kind of retrofit. For example, adding R-15 of insulation to existing insulation in the attic would be a "relative" retrofit parameter value, whereas a window replacement would consist of several "absolute" retrofit parameter values. This distinction becomes important when applying retrofit measures to the calibrated base case because key base case parameter values are considered uncertain. The test designer uses the nominal retrofit models to generate post-retrofit energy consumption results and savings.

The test designer introduces input uncertainty into the pre-retrofit test specification (nominal model). This represents the uncertainty associated with collecting audit survey data and developing inputs from that data. This step includes:

- Using the nominal pre-retrofit model to perform sensitivity tests on inputs with potentially high uncertainties to determine their relative effects on outputs, and selecting the inputs that have both substantial uncertainties and effects on outputs as approximate inputs.
- Specifying an uncertainty range (approximate input range) for each approximate input.

- Selecting explicit input (surrogate truth) values from within the approximate input ranges. The selection can either be done through a random process or through a manual process to create test cases with specific characteristics. It is useful to choose combinations of explicit input values that yield high, near nominal, and low pre-retrofit energy use relative to the nominal model. Those who will be performing the calibrations, the calibrators, must be blind to the explicit inputs.

Box R1 in Figure 1

The test designer performs simulations using the explicit inputs to generate the pre-retrofit (base case) synthetic utility bill data. This is typically monthly electric and gas consumption data, but the method can be used to generate and test against higher frequency or lower frequency synthetic building energy performance data. Also, end-use data at varying levels of disaggregation can be used, mimicking the availability of sub-metered data.

Box R2 in Figure 1

The test designer performs simulations to generate explicit (synthetic truth) post-retrofit energy consumption and savings results. This involves starting with the appropriate explicit pre-retrofit base-case model, and adjusting appropriate base case inputs for each retrofit case and combinations of cases.

Box C1, C2, and C3 in Figure 1

The calibrator (test taker) generates results using the calibration technique being tested and its associated simulation program. The associated simulation program must be the same simulation program used by the test designer. The calibrator may only know permissible data as defined by the test designer. Typically, this would mean that the calibrator could know the nominal model inputs, the "synthetic" pre retrofit utility bill consumption data, and the approximate input ranges, but could not know the explicit inputs or the explicit post retrofit outputs and savings results. This kind of test imitates what an auditor or modeler would typically know when trying to model and calibrate to a real building. An audit would reveal a set of probable input values analogous to the nominal inputs in the test. An experienced auditor or modeler would have an idea of the approximate uncertainty associated with the most influential input values analogous to the approximate input ranges for the test. The auditor or modeler would use a year or more of utility bills to perform the calibration analogous to the synthetic utility consumption data furnished by the test designer. The calibrator first runs the nominal (uncalibrated model) to generate the pre and post retrofit uncalibrated results (consumption data) (Box C1) and then creates and runs the calibrated model to generate the pre and post retrofit calibrated results (consumption data) (Boxes C2 and C3).

The calibrator predicts energy savings via one of the following calibration approaches:

- Calibrate the base-case model inputs using the synthetic utility bills, then apply the specified retrofit cases to the calibrated model.
- Apply the specified retrofits to the uncalibrated base-case model and then calibrate or correct energy savings predictions using the synthetic utility bills (without adjustment to base-case model inputs), e.g., as (calibrated savings) = (predicted savings) × (base case actual bills)/(base case predicted bills).
- Other calibration methods. This test method makes no recommendation about how to perform calibrations. Any calibration method that seeks to improve energy savings predictions through use of pre-retrofit building energy performance data may be tested via this method.

The test designer and calibrator use, at least, the following comparisons (figures of merit) to evaluate the adequacy of the tested calibration technique:

- Compare the goodness of fit between the calibrated energy consumption data and the explicit (synthetic truth) energy consumption data for the pre and post-retrofit cases.
- Compare the calibrated savings predictions to the explicit (synthetic truth) savings.
- For programs where a calibrated pre retrofit (base case) model is applied, compare calibrated input values to the corresponding explicit (truth) input values.

All of these comparisons are important for assessing the accuracy of the calibration method. A large disagreement in any one of them indicates the presence of compensating errors or some other error. Not all calibration methods will allow all of the above comparisons, however, all calibration methods will allow comparison of the savings predictions from the tested simulation tool and any associated calibration techniques, to the savings predictions from the same tool run with the explicit (synthetic truth) inputs. The calibrated results and savings may also be compared to the nominal (uncalibrated) results and savings to assess the benefit of calibration (Judkoff et. al., 2010, Appendix G).

The method for testing model calibration techniques described above is a “pure” calibration test in that the synthetic utility billing data is generated with the tested program, and the program accuracy related to building physics modeling is not tested. As noted previously, the pure calibration test requires strict adherence to the restriction that the calibrator is limited to permissible data until the testing is completed. This method facilitates “self-testing” of a calibration technique, and is useful in several ways, including: a) testing a single calibration method to see how well it works under a variety of test conditions, b) testing several calibration methods to determine under what test conditions each is best, c) investigating how much, and what kind, of informational content is needed in the synthetic pre-retrofit data to achieve good calibrations with different calibration methods (eg. monthly vs daily vs hourly data and availability of different types of submetered or disaggregated data), d) testing with various amounts and kinds of “noise” in the synthetic data, and e) diagnostic testing.

The pure calibration test may not be practical for a certification test that must be administered by a third party organization and where an honor system is not deemed appropriate. A method to facilitate third party testing which assures that the person performing the test does not know the explicit inputs, has also been developed (Judkoff et al. [2011a, 2011b]) and is referred to as the “reference program method.” The main difference between the two test methods is that for the reference program method several (preferably at least three) reference programs are used to generate the synthetic utility bills, and to create the reference energy savings data. The bills and the savings are taken as the average of the reference program results. The reference program method is both a test of the calibration technique, and a test of how closely the physics models in the tested program match the physics models in the reference programs. Example acceptance criteria may be used to facilitate the comparison of energy savings predictions (Judkoff et al 2011b).

The test method is defined at a high level to allow many different kinds of testing of calibration techniques. Here we have focused on the most typical case involving utility bill consumption data however, in a research context calibration might be conducted using temperature data, measurements from heat flux transducers, or any other energy related measurements. For typical situations BESTEST-EX provides specifications and test cases that can be used to avoid the work of creating tests and input specifications. Robertson, Polly, and Collis (2013) used a combination of the BESTEST-EX specifications and their own

specifications. They also compared different calibration techniques using monthly, daily, and hourly synthetic energy use data. The method has also been applied at Oak Ridge National laboratory (New 2016). The method accommodates the creation of highly diagnostic test sequences. For example, the following conceptual sequence has been proposed (Judkoff 2014):

- Create a test with one uncertain influential input parameter. If a calibration technique can't close on the "truth" value for one parameter, it is seriously flawed. If the calibration technique compares well on test 1, then,
- Create a test with two uncertain influential uncorrelated input parameters. Then,
- Create a test with two uncertain influential correlated input parameters. Then,
- Create a test with multiple uncertain influential input parameters with zero infiltration, adiabatic floor or "floating floor", and an extended time (one month) in the weather file with no sun and constant cold outdoor dry bulb temperatures. Given this information, a human with knowledge of building physics could determine the overall heat transmission coefficient of the building. Can the calibration technique pick up this signal in the data such that the sum of the individual calibrated parallel heat transfer components of the building envelope match the "true" overall heat transmission coefficient? This is equivalent to throwing a physics "softball" at the calibration technique. Then,
- Etc.

Conclusions

An analytical method to test model calibration techniques has been created and field tested. The method is broadly defined at a high level to accommodate many different kinds of tests of calibration techniques ranging from simple and highly diagnostic to complex and relatively realistic. The method can be used in many ways including but not limited to:

- Self-testing of a calibration method and associated simulation program to see how well it works under a variety of test conditions.
- Testing of several calibration methods to determine under what test conditions each is best
- Testing of the informational content needed in the synthetic utility bill energy consumption data.
    - testing using monthly, daily, or hourly data
    - testing assuming the availability of various amounts and types of sub-metered data
- Testing with various amounts and kinds of noise in the "synthetic truth" data.
- Diagnostic testing.

The test method is in the process of being adapted as a Standard Method of Test.

References

ANSI/ASHRAE Standard 140-2014, *Standard Method of Test for the Evaluation of Building Energy Analysis Computer Programs*

ANSI/BPI Standard 2400-S-2015, *Standard Practice for Standardized Qualification of Whole-House Energy Savings Predictions by Calibration to Energy Use History.*

ASHRAE Guideline 14-2014, *Measurement of Energy and Demand Savings.*

Judkoff, R. 2014, Presentation to the RESNET Calibration Standard Method of Test Working Group (slide 14). February 23, 2014. National Renewable Energy Laboratory (NREL).

Judkoff, R., and J. Neymark. 1995a. *International Energy Agency Building Energy Simulation Test (BESTEST) and Diagnostic Method.* National Renewable Energy Laboratory (NREL) Publications Database www.nrel.gov.

Judkoff, R., and J. Neymark. 1995b. *Home Energy Rating System Building Energy Simulation Test (HERS BESTEST), Vol. 1: Tier 1 and Tier 2 Tests User's Manual and Vol. 2: Tier 1 and Tier 2 Tests Reference Results.* National Renewable Energy Laboratory (NREL) Publications Database www.nrel.gov.

Judkoff, R., B. Polly, M. Bianchi, J. Neymark. 2010. Building Energy Simulation Test for Existing Homes (BESTEST-EX). NREL/TP-550-47427. Golden, CO, USA: National Renewable Energy Laboratory. http://www.nrel.gov/docs/fy10osti/47427.pdf

Judkoff, R., J. Neymark, B. Polly, M. Bianchi. 2011a. "The Building Energy Simulation Test For Existing Homes (BESTEST-EX) Methodology." Proceedings Building Simulation 2011a, International Building Performance Simulation Association. Also see pre-printed version, NREL CP-5500-51655. National Renewable Energy Laboratory, Golden, CO. http://www.nrel.gov/docs/fy12osti/51655.pdf.

Judkoff, R., B. Polly, M. Bianchi, J. Neymark, M. Kennedy. 2011b. Building Energy Simulation Test for Existing Homes (BESTEST-EX): Instructions for Implementing the Test Procedure, Calibration Test Reference Results, and Example Acceptance-Range Criteria. NREL TP-5500-52414. National Renewable Energy Laboratory, Golden, CO. http://www.nrel.gov/docs/fy11osti/52414.pdf.

New, J. 2016. *Autotune Calibration and Trinity Test Evaluation.* ASHRAE Seminar #59, Winter Technical Program, Orlando FL. January 27 2016.

Neymark, J., and R. Judkoff. 2002. *International Energy Agency Building Energy Simulation Test and Diagnostic Method for Heating, Ventilating, and Air-Conditioning Equipment Models (HVAC BESTEST); Volume 1: Cases E100-E200.* National Renewable Energy Laboratory (NREL) Publications Database www.nrel.gov.

Neymark, J., and R. Judkoff. 2004. *International Energy Agency Building Energy Simulation Test and Diagnostic Method for Heating, Ventilating, and Air-Conditioning Equipment Models (HVAC BESTEST): Volume 2: Cases E300-E545.* National Renewable Energy Laboratory (NREL) Publications Database www.nrel.gov.

Neymark, J., R. Judkoff, I. Beausoleil-Morrison, A. Ben-Nakhi, M. Crowley, M. Deru, R. Henninger, H. Ribberink, J. Thornton, A. Wijsman, and M. Witte. 2008a. *International Energy Agency Building Energy Simulation Test and Diagnostic Method (IEA BESTEST): In-Depth Diagnostic Cases for Ground Coupled Heat Transfer Related to Slab-on-Grade Construction.* National Renewable Energy Laboratory (NREL) Publications Database www.nrel.gov.

Neymark, J., R. Judkoff, D. Alexander, C. Felsmann, P. Strachan, A. Wijsman. 2008b. *International Energy Agency Building Energy Simulation Test and Diagnostic Method (IEA BESTEST) Multi-Zone Non-Airflow In-Depth Diagnostic Cases: MZ320 -- MZ360.* National Renewable Energy Laboratory (NREL) Publications Database www.nrel.gov.

Reddy, T.A., I. Maor, C. Panjapornporn, and J. Sun. 2006. *Procedures for reconciling computer-calculated results with measured energy use, RP-1051 final research report*. Atlanta: American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc.

Robertson, J.; Polly, B.; Collis, J. (2013). *Evaluation of Automated Model Calibration Techniques for Residential Building Energy Simulation*. 91 pp.; NREL Report No. TP-5500-60127.