Suitability of ASHRAE Guideline 14 Metrics for Calibration

Aaron Garrett, PhD

Joshua R. New, PhD
Member ASHRAE

ABSTRACT

We introduce and provide results from a rigorous, scientific testing methodology that allows pure building model calibration systems to be compared fairly to traditional output error (e.g. how well does simulation output match utility bills?) as well as input-side error (e.g. how well, variable-by-variable, did the calibration capture the true building's description?). This system is then used to generate data for a correlation study of output and input error measures that validates CV(RMSE) and NMBE metrics put forth by ASHRAE Guideline 14 and suggests possible alternatives.

INTRODUCTION

In previous work (New 2012, Garrett 2013, Garrett 2015), the Autotune calibration system was used to calibrate a Building Energy Model (BEM) to a highly instrumented and automated ZEBRAlliance research home (Biswas 2011) by fitting measured monthly load and electrical data. This research showed that the evolutionary computation approach to automatic calibration was effective in fitting the EnergyPlus output from the calibrated model to the measured data. This calibration included time-varying parameters such as occupancy and equipment schedules necessary for practical application. There are other detailed studies which have compiled approaches to calibration (Reddy 2006) and the performance of many calibration methods (Coakley 2014).

However, because the tuning was applied to a real building with unknown model parameters (thus the need for calibration), it was impossible to determine exactly how well the tuned model matched the actual building in terms of model parameters over the course of a year. Even with costly lab-controlled research homes involving perfectly repeated automated occupancy, measurement of materials entering the building, and documentation of the construction process, it is still impractical to track the exact value of physical parameters for all materials throughout the building as they change with time.

In this work, rather than attempting to calibrate existing buildings to match measured data, we instead attempt to calibrate fully-specified Department of Energy commercial reference buildings to match EnergyPlus output generated from altered versions of those buildings (where the altered model parameters are known to calibration test designers but unknown to calibrators) using the pure calibration technique described in BESTEST-EX (Judkoff 2011). This allows one to test a calibration process's accuracy on both model outputs and model inputs under ideal laboratory conditions, providing a true benchmark for comparing all calibration methods. With a centralized benchmarking system for BEM calibration, it becomes possible to compile performance metrics from calibration algorithms applied to (suites of) calibration problems to allow certification, rating, or selection of a calibration process that typically performs best for a given, real-world calibration problem

and accompanying data.

Trinity Testing

On July 16, 1945, the United States tested the detonation of the first nuclear weapon ever created. One of the chief scientists involved in the project, J. Robert Oppenheimer, named the test "Trinity" which was designed to fully determine the efficacy of any nuclear explosive device. In this work, the name "Trinity" is adopted as a convenient term to refer to the implemented testing framework that can determine the effectiveness of any (automatic or manual) building model calibration system which quantifies the accuracy in terms of input-side error metrics. The calibration technique underlying the "Trinity test" system was first developed and named the "pure calibration test method" by Ron Judkoff from the National Renewable Energy Laboratory (NREL) as part of BESTEST-EX (Judkoff 2011). The "Trinity test"' name was first used in relation to BEM calibration by Amir Roth, the Department of Energy's (DOE) Building Technologies Office technology manager overseeing this project. The Trinity test system has been deployed for public use at http://bit.ly/trinity_test.

The Trinity test framework is designed to deal with common issues inherent in auto-calibration results:

- Most calibrations in the literature are carried out on specific, unique buildings of interest. Building data often is not shared, which complicates any attempt by other investigators to duplicate the work.
- Researchers often report the results of their calibrations in different ways using different metrics, and nearly all results detail only the model output. If a real building is used, then exact components of the building are likely unknown, which is precisely why automatic calibration is needed. This leads to a proliferation in the literature of necessarily unique, largely irreplicable, and essentially incomparable results that do not help mature the state of the art in automatic calibration approaches.

A solution to all of these problems is to test calibration approaches using modified benchmark models. For instance, a given Department of Energy commercial reference building has a fully specified EnergyPlus model, which produces noise-free output (e.g. no real-world calibration drift or other uncertainty for measured data, and no gap between the simulation algorithms versus real-world physics) when passed through EnergyPlus. Using such a model as a base model, a controlled test model can be created where certain variables of the base are modified within some specified bounds (e.g., within ±30% of the base value). By selecting a valid value for each of the input parameters, a test creator can define what we refer to as the "true model". The true model can be passed through EnergyPlus to produce similar noise-free output which functions as a surrogate for clean sensor data. Then, anyone interested in testing a calibration approach can simply retrieve the base model, including names and ranges of the modified variables, and the true model's EnergyPlus output.

Ideally, a calibration procedure would be able to discover the (hidden) input variable values of the true model in addition to producing very similar EnergyPlus output with the calibrated model. Thus, the calibration system's effectiveness can then be measured exactly by its error in the input domain (true vs. calibrated variable values) and output domain (true vs. calibrated model EnergyPlus output). In the context of the Trinity test system, we use "calibration" primarily in comparison to output of a reference simulation as a surrogate to measured data, which varies slightly from ASHRAE Guideline 14 definition (b) of calibration "process of reducing the uncertainty of a model by comparing the predicted output of the model under a specific set of conditions *to the actual measured data* for the same set of conditions."

The Trinity test system is illustrated in Figure 1. Here, the base model and true output are given to the Calibrator, while the true model is maintained privately by the Evaluator. The Calibrator would benefit from having the names and valid ranges of the variables which should be calibrated, and these elements are provided as separate files or within an XML version of an EnergyPlus input file. The Evaluator, by having the actual, fully-specified, true model,

can assess not only the accuracy of the predicted model output, but it can also assess the accuracy of the predicted model's variables to the true, hidden variables.

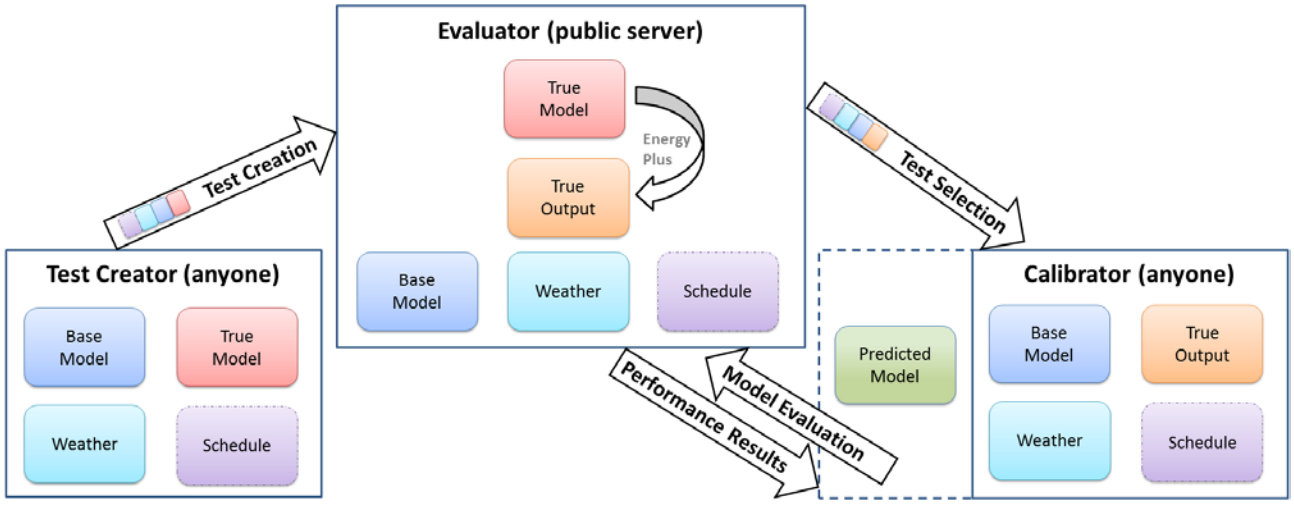

Figure 1. Trinity Test System. The Test Creator produces a base model, weather file, and schedule (optional) along with a model that constitutes the "true" model. This true model can be automatically generated and has a setting for each input parameter that obeys the ranges, distributions, and mathematical constrains specified for tunable parameters in the base model. These files are given to the Evaluator when the test is created, at which time the Evaluator produces the true output generated by the true model. The Calibrator selects a test from the Evaluator and receives the base model, weather, schedule, and true output. The Calibrator then performs the tuning to produce a predicted model, which is submitted to the Evaluator for evaluation. The Evaluator compares the predicted model to the true model, both in input and output, and returns the results in the form of aggregated statistics that quantify the Calibrator's accuracy at recovering the true model.

Trinity Test Limitations. It should be pointed out that while Trinity testing is a scalable, automated methodology that does not require extensive manual labor for measuring and maintaining a real-world experimental facility, it has many limitations – some of which we discuss below.

First, the use of a simulation alone (EnergyPlus in this case) means all noise has been removed from the system. Practical issues such as sensor drift, model-form uncertainty due to algorithmic inaccuracies, missing values from utility bills (or sensors), different dates of utility bill measurements, and other complications faced by a calibration methodology employed in practical use are not captured in this testing methodology. The Trinity test could be extended to allow systematic exploration of calibration inaccuracies caused by such normal phenomena by systematically adding noise and missing values to EnergyPlus output prior to calibration. To identify model-form uncertainty inaccuracies that arise in differences between a simulation engine's algorithms and real-world physics, the interested reader is referred to the BESTEST-EX building physics test suites (Judkoff 2011).

Second, a specific weather file is used to define environmental conditions. While our implementation allows a test creator to provide the weather data, this is most frequently Typical Meteorology Year (TMY) data. For real-world application, Actual Meteorological year (AMY) weather data should be used for the time period during which utility bills and/or sensor data was collected. Previous research has shown that annual energy consumption can vary by ±7% and monthly building loads by ±40% based solely on which weather vendor provides the AMY data (Bhandari2012).

Third, there is not always an intuitive mapping between a point-measurement in a real building and an EnergyPlus output. As an example, a wall in EnergyPlus reports its average temperature but stratification of thermal gradients in an actual building would require either a precise sensor location or processing of a series of temperatures to correspond with what EnergyPlus reports as the interior or exterior temperature of a given wall. Optionally, a small

sub-wall at the sensor location can be created in the EnergyPlus model, as the authors used in the ZEBRAlliance calibration, but this drives up runtime and is not currently automated through use of sensor location data.

Fourth, real-world measurements are best taken with NIST-calibrated sensors in ways that adhere to known standards. As an example, heat flux measurements from a surface are usually taken according to ASTM E2684, which requires an appropriate sensor to be measured under a thin layer of the material to ensure the presence of the sensor doesn't disrupt the temperature and continuity of the material. As of the time of this writing, a new feature is being considered for EnergyPlus that would allow it to be compared in validation studies where heat flux transducers are throughout a multi-layer envelope assembly (e.g., a wall). Exquisite care in measurement, as well as extensions to the simulation engine itself, is often necessary to allow a proper comparison.

Fifth, all inputs are treated equally and aggregate metrics (to limit gaming) are provided for input-side error. Trinity testing does not incorporate domain-specific information that some inputs matter more than others when it comes to the effect on whole-building energy consumption or given model use case (e.g. optimal retrofit). Being a clean-room methodology which does not address real-world complications (sensor drift, missing measured data, lack of measurement correspondence with simulation output, inaccurate algorithms, etc.) it does not necessarily follow that a calibration methodology which performs well under Trinity test conditions is field-deployable.

Automated Web Service

The Trinity testing framework, as presented above, requires a strict protocol. The Calibrator must never be exposed to the true model, which should remain private along with any information that might be used to infer the values of those hidden variables. Therefore, the Calibrator is entirely dependent upon the Evaluator to assess the accuracy of the calibration process (since only the evaluator has access to the true model). This requires a great deal of effort from the Evaluator, which is a limitation of the methodology. To alleviate this, the Trinity testing framework has been automated by converting it into a web service.

The functionality of the Trinity service consists of four actions:

1. Create a test case
2. Download a test case
3. Evaluate a calibrated model
4. Download evaluation results

A user would invoke their calibration procedure—which could be manual, semi-automatic, or fully automated—between steps 2 and 3. Each of these four Trinity test actions in the workflow is described in more detail in the following subsections.

Test Creation. A test case is created when a user (which we refer to as the test creator) uploads to the service: a base model (which includes tunable parameter ranges), weather, optional schedule, and the true model that was derived from the base. Functionality exists for automatically selecting a true model from a customizable range (e.g. ±30%) from the base value, but is not exposed to the web service. Only the Test Creator has permission to access the true model. Once a test case has been submitted, the service processes the true model using EnergyPlus to produce the noise-free sensor data output that will be used for calibration. Since this process involves running an EnergyPlus simulation that can be time-consuming by website/service standards, depending on the model, submitted tests are queued by the system and processed in a first-come, first-served fashion. A current design constraint we imposed on the current web service is that the authors needed to allow comparison to previous calibration procedures, and thus the web service only supports EnergyPlus 7.0 at this time

Test Selection. After a test has been created, Calibrators may select individual test cases against which to assess themselves. Once a test is selected, the Calibrator receives the base model, any supplemental files (weather, schedule, etc.), and the system-processed sensor data output from the true model. The Calibrator also receives information

detailing the variables that should be calibrated as a part of the base model (as detailed in the next section).

Model Evaluation. After a Calibrator has processed a particular test case and arrived at a calibrated (predicted) model, the Calibrator can then submit the model to the Trinity system in order to have it evaluated against the true model. Once again, since this requires running an EnergyPlus simulation, this is a potentially time-consuming process so the system queues such requests in a first-come, first served fashion in order to calculate the results of the assessment.

Results Extraction. Once calibration results are available, the Calibrator may download the full results of the assessment. These results include the industry-standard CV(RMSE) and NMBE measures for every individual output field (e.g., total building electrical load), as well as aggregate errors for input variables, defined as percentages of the specified range. This means that Calibrators will not be given actual per-variable errors for input variables. Doing so would provide additional information about the hidden values, which might allow clever Calibrators to increase their performance on the test artificially. Instead, the percentage errors for each variable are aggregated, and their minimum, maximum, average, and variance are reported. For instance, a Calibrator might receive an assessment reporting that the calibrated variables had an average error of 22% (variance 17%) with a range from 4% to 57%.

Providing Supplemental Information Using XML

As specified in the previous section, ideally the only files required from the test creator would be the base model, true model, weather, and schedule (if used). However, in practice, a Calibrator requires some amount of supplemental information detailing which variables should be calibrated and through what ranges the variable values extend. Rather than requiring the Test Creator to produce and distribute additional files, the Trinity service allows model files to contain this additional information by representing them using eXtensible Markup Language (XML).

The XML format is more expressive than the default EnergyPlus IDF format because systems can simply ignore unrecognized and unused tags and attributes. For the Trinity service, a web-based conversion system was created to translate EnergyPlus models in IDF format into an equivalent XML format (and vice versa).

A sample of the basic XML format is shown in Listing 1.

```
<Material>
    <Name>
        Metal Roofing
    </Name>
    <Roughness>
        MediumSmooth
    </Roughness>
    <Thickness>
        0.0015
    </Thickness>
    <Conductivity>
        45.0060
    </Conductivity>
```

Listing 1. This is an excerpt from a sample XML building model.

Using this XML format, attributes can be added to tags that inform the Calibrator that a given input variable is to be calibrated, as well as the allowable range of that variable. For instance, if the `Thickness` variable were calibrated from the model in Listing 1, its XML tag might be modified to be more like Listing 2.

```
<Material>
    <Name>
        Metal Roofing
    </Name>
    <Roughness>
        MediumSmooth
    </Roughness>
    <Thickness tuneType="float" tuneMin="0.0010" tuneMax="0.0030"
               tuneDistribution="gaussian" tuneGroup="G07"
               tuneConstraint="G05+G06+G07 &lt; 1">
        0.0015
    </Thickness>
    <Conductivity>
        45.0060
    </Conductivity>
```

Listing 2. This is the fully-specified version of the XML model from Listing 1.

Here, the `tuneType`, `tuneMin`, and `tuneMax` attributes have been added to the `Thickness` tag in order to specify both that this variable should be calibrated and that it should be treated as a continuous-valued variable with an allowable range of [0.0010, 0.0030]. The web-based conversion utility can accept such a modified XML model file and return the corresponding EnergyPlus IDF file, along with a standard CSV file containing all of the variables that have been marked to be calibrated (along with their ranges). Additional functionality that can take an IDF and CSV file to create an XML file with appropriate markup has been generated, but is not currently available through the web service.

In addition to the allowed calibration range, the Test Creator may specify any number of additional parameters that would be useful for the Calibrator. Listing 2 illustrates the parameters that the current XML-IDF conversion system can extract. These parameters include the following:

- required `tuneType` that defines the type of variable (`float` or `integer`)
- required range (`tuneMin` and `tuneMax`) in which the variable's calibrated value should be found
- optional statistical distribution `tuneDistribution` which this variable's allowable values should obey. This could be extended to support the 200+ distributions supported by Python's SciPy package.
- optional group name `tuneGroup` that allows multiple variables to be calibrated as one entity (such that all variables in the same group have the same calibrated value)
- optional constraint equation `tuneConstraint` that defines how variables interact with one another

The constraint equation, if used, requires that all variables associated with the constraint be assigned group names. In the example in Listing 2, the constraint equation specifies that the sum of the calibrated values for groups `G05`, `G06`, and `G07` must be less than[1] 1. It is important for Calibrators to be aware of such constraints because it is possible for calibrated values individually to be within their ranges but collectively to violate the constraint(s).

For instance, suppose an EnergyPlus model contains calibration variables for the return air fraction, fraction radiant, and fraction visible for a given light in the model. Also suppose each variable is a percentage with an allowable range in [0.3, 0.7]. In this case, EnergyPlus calculates the fraction of the heat from the lights convected to the zone air as

[1] Note that the valid XML code for the less-than symbol (<) must be used here because the less-than and greater-than symbols are meaningful XML identifiers.

$$F_{convected} = 1.0 - (F_{returnAir} + F_{radiant} + F_{visible}).$$

In other words, the three variables under consideration must have a sum that is no greater than 1.0. It is possible for the variables to be individually within the allowed ranges (say 0.5, 0.4, and 0.6) but for the set to violate the underlying EnergyPlus simulation (total = 1.5). If a calibration technique searches through the space of *individually* feasible values, it may generate a solution that is *collectively* infeasible. Specification of the constraints allows the Calibrator to avoid generating such infeasible solutions.

This has the effect of pruning the search space for a calibration process and limits time running simulations that will cause EnergyPlus to crash for cases in which a clearly-defined rule avoids such system states. There is currently no consolidated location for these rules, but the most common rule recommended is that cooling and heating setpoints typically should not overlap.

METHODS

In a previous report (Garrett2013b), the authors investigated the use of a Trinity-like test approach to determine the efficacy and efficiency of Autotune on three Department of Energy commercial reference buildings (Medium Office, Stand-alone Retail, and Warehouse). The results of that initial investigation were very promising and warranted further study. The Autotune system has thus far produced very accurate tuned models in terms of output error, CV(RMSE)<4% with a fully automated process. However, the mean input error for most reference buildings is above 20%. In an effort to reduce input-side error as the primary metric, the authors asked the question whether the CV(RMSE) and NMBE metrics codified by ASHRAE Guideline 14 (ASHRAE 2002) are even the appropriate ones to use. Here, we investigate whether there exist any correlations between measures of output error and measures of input error. The hypothesis is that, if such correlations were high enough, using the appropriate output error measures during tuning would correspondingly decrease the input errors.

Selecting the Test Models

For this work, the authors chose to use 15 of the 16 commercial Department of Energy reference buildings (Deru 2011), which represent about 70% of all commercial buildings in the U.S. The Outpatient reference building was excluded. In all cases, the models were for new constructions located in climate zone 5A (Chicago, Illinois).

Determining the Calibration Variables

The authors consulted domain experts for each building to identify the most common and effective variables to calibrate. These variables differ from building to building. Table 1 provides a high-level summary for the number of each building's variables. A complete description of each of the 1,810 input file parameters is too lengthy to disclose in this publication, but is available as supplemental material. In Table 1, the number of inputs and number of groups are listed for each building. The number of inputs is simply the number of building variables that were tuned. However, some variables belong to groups that should all share the same value. The number of groups displays the true number of degrees of freedom available to the calibration system. Table 1 is sorted in ascending order by number of groups, as this represents a type of proxy for calibration difficulty. In order to determine the correlations between input and output error measures, many model files must be generated for each reference building type. In this work, 1,000 models were randomly generated for each reference type, where the calibration variable values were randomly selected from their allowable ranges.

Specifying the Input and Output Metrics

Each generated model was compared to the base reference building to determine the input and output errors. The measures of output errors were CV(RMSE), RMSE, NMBE, MBE, MAPE, correlation, and kurtosis. The measures of input errors were CV(RMSE), RMSE, NMBE, MBE, MAPE, and percent absolute error (PAE).

Table 1. Calibration variables used for each reference building, in ascending order by calibration difficulty based on the number of groups used.

Building	Inputs	Groups
Medium Office	81	36
Large Office	85	43
Warehouse	47	44
Full Service Restaurant	49	49
Quick Service Restaurant	54	54
Stand-alone Retail	59	55
Small Office	72	58
Large Hotel	110	67
Super Market	78	72
Midrise Apartment	155	78
Strip Mall	113	85
Primary School	166	109
Secondary School	231	122
Small Hotel	282	131
Hospital	227	139

RESULTS

In order to perform the analysis, the correlations between every input and output measure must be calculated for each building type. Figure 2 displays the correlations between each input and output error measure for each building for the |Electricity:Facility [J](Hourly)| output variable, which was the one that was used for tuning in (Garrett2013b). As is clear from the figure, there is no strong correlation between any pair of input and output error measures for any building.

However, perhaps a different output variable may show stronger correlation and would be better suited to be incorporated into the tuning. To determine this, the authors considered only the top five most highly correlated output variables. In this case, the RMSE and MBE were eliminated from consideration, since their correlation is identical to CV(RMSE) and NMBE due to their nearly identical mathematical definition. Figure 2 shows that two output variables in particular, |InteriorEquipment:Electricity [J](Hourly)| and |InteriorLights:Electricity [J](Hourly)| typically show very strong correlation between measures of input and output error.

To ensure that these variables are consistently correlated across building types, we count the number of unique buildings which appear in the top five correlations as shown in Table 2. The mean correlation for |InteriorEquipment:Electricity [J](Hourly)| was 0.4795935. Likewise, for |InteriorLights:Electricity [J](Hourly)| the mean was 0.4564204. The practical implication of this finding is that ASHRAE Guideline 14 could recommend CV(RMSE) and NMBE calibration requirements for tier-2 metrics (beyond whole-building electrical) and these likely should include interior electrical equipment, interior lights, and HVAC energy use.

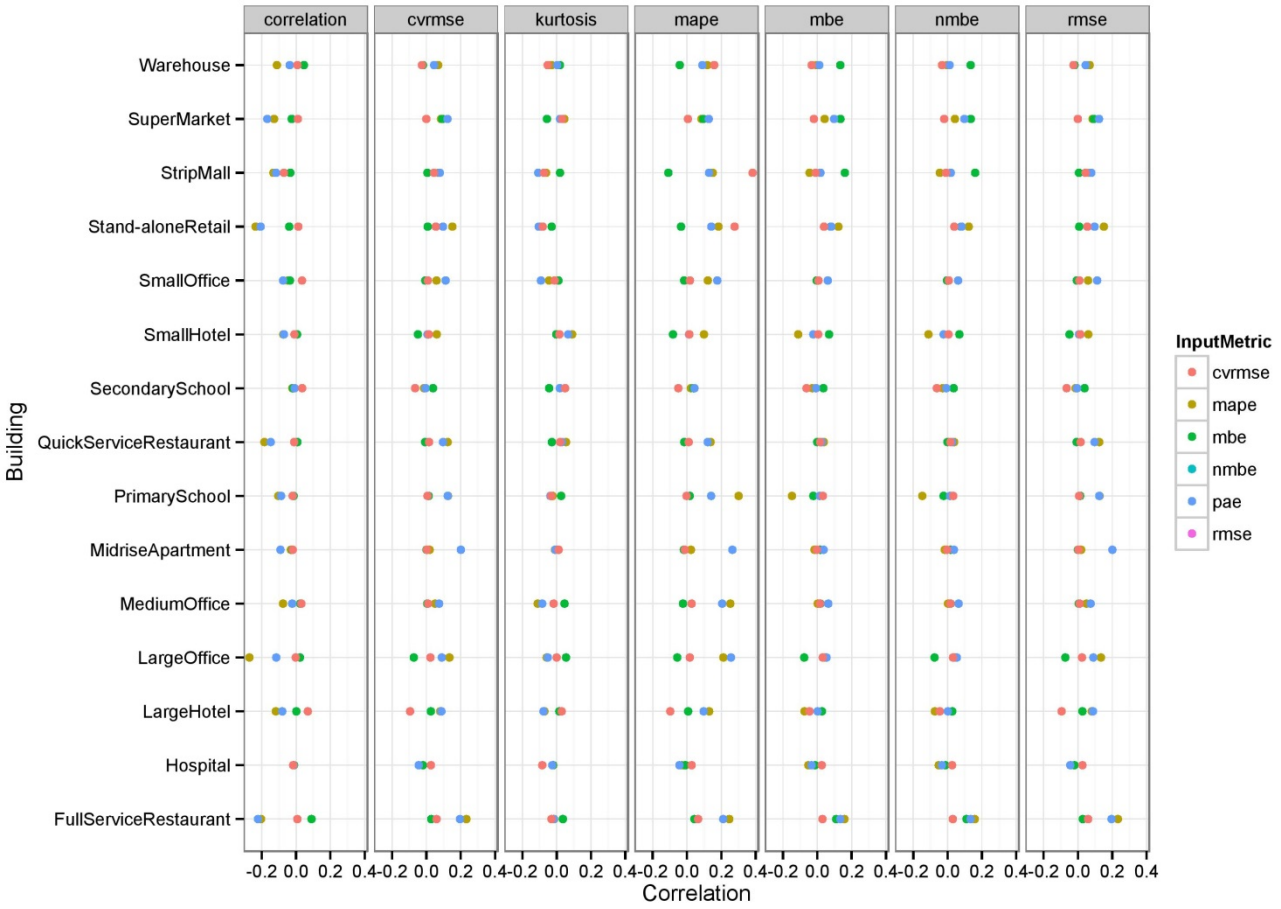

Figure 2. The correlations between input and output metrics for hourly facility electricity usage are represented as dot-plots. The lack of any statistically significant difference shows that the CV(RMSE) and NMBE metrics used in ASHRAE Guideline 14 are as good as any of those tested regarding correlation to input-side error.

Table 2. The number of unique buildings in the top five correlations

	Output Variable	Number of Buildings
1	Electricity:Facility [J](Hourly)	2
2	Gas:Facility [J](Hourly)	3
3	Gas:Facility [J](Monthly)	1
4	Heating:Gas [J](Hourly)	1
5	InteriorEquipment:Electricity [J](Hourly)	10
6	InteriorEquipment:Gas [J](Hourly)	2
7	InteriorLights:Electricity [J](Hourly)	9
8	PSZ-AC_1:1:Air Loop Total Heating Energy[J](Hourly)	1
9	Water Heater:WaterSystems:Gas [J](Hourly)	1
10	COOLSYS1:Plant Loop Unmet Demand [W](Hourly)	1
11	SWHSYS1:Plant Loop OutletNode Temperature [C](Hourly)	1
12	SWHSYS1:Plant Loop Unmet Demand [W](Hourly)	1
13	PSZ-AC:2:Air Loop Total Heating Energy[J](Hourly)	1
14	PSZ-AC:2:AirLoopHVAC Outdoor Air Economizer Status(Hourly)	1
15	PSZ-AC:3:Air Loop Total Heating Energy[J](Hourly)	1
16	PSZ-AC:4:Air Loop Total Heating Energy[J](Hourly)	1
17	FURNACE_PACU_CAV_1:1:Air Loop System Cycle On/Off Status(Hourly)	1
18	FURNACE_PACU_CAV_1:1:Air Loop Total Cooling Energy[J](Hourly)	1
19	FURNACE_PACU_CAV_1:1:Air Loop Total Heating Energy[J](Hourly)	1

CONCLUSIONS

The Trinity test system for EnergyPlus, based on the pure calibration technique of ASHRAE BESTEST-EX (Judkoff 2011), has been formalized and deployed for free use by industry at http://bit.ly/trinity_test. The test system and supporting web service allow anyone to define their own publicly-available calibration problems and automatically provides performance metrics for calibration on any of those tests. The system not only reports the output-side error metrics currently used by codes and industry, but also establishes the importance of input-side error (how faithfully the algorithm recovers the true building parameters) as the primary performance metric by which a calibration algorithm should be judged. The test system enables the identification of a calibration algorithm that performs best on the uncertainty parameters and measured data available for a given, real-world calibration scenario.

In an effort to improve the state-of-the-art in calibration algorithms, the authors have called into question the use of the CV(RMSE) and NMBE metrics canonized by ASHRAE Guideline 14. If other metrics for output-side error showed a high correlation to input-side error testable via BESTEST-EX, then it should be proposed as alternative metrics in ASHRAE Guideilne 14. The authors conducted a study involving over 20,000 calibrations for 15 DOE reference buildings each with 36-139 inputs calibrated such that simulation output matched whole-building electricity. Unfortunately, the correlations between input and output error measures were not statistically significant, implying that the metrics put forth in ASHRAE Guideline 14 are as good as any of the 4 other binary metrics tested. However, analysis revealed that the hourly interior equipment and lighting electricity correlations were reasonably high (mean of over 0.45 across all reference buildings). Further work should investigate whether tuning based on these metrics in addition to or instead of hourly facility electricity produces lower input error. In addition, these important input variables could be weighted proportional to their importance, as informed by uncertainty quantification studies, in order to realize a lower input-side error.

ACKNOWLEDGEMENTS

Funding for this project was provided by field work proposal CEBT105 under the Department of Energy Building Technology Activity Number BT0201000. The authors would like to thank Zheng O'Neill for identifying the input parameters, ranges, and distributions used in this study.

REFERENCES

ASHRAE (2002). Measurement of Energy and Demand Savings. ASHRAE Guideline 14-2002. Available: https://gaia.lbl.gov/people/ryin/public/Ashrae_guideline14-2002_Measurement%20of%20Energy%20and%20Demand%20Saving%20.pdf

Bhandari, Mahabir S., Shrestha, Som S., and New, Joshua R. (2012). Evaluation of Weather Data for Building Energy Simulations. In Journal of Energy and Buildings, volume 49, issue 0, pp. 109-118, June 2012

Biswas, K., Gehl, A., Jackson, R., Boudreaux, P., and Christian, J. Comparison of Two High-Performance Energy Efficient Homes: Annual Performance Report, December 1, 2010 – November 30, 2011. Oak Ridge National Laboratory report ORNL/TM-2011/539. Available: http://info.ornl.gov/sites/publications/files/Pub34163.pdf

Daniel Coakley, Paul Raftery, Marcus Keane (2014). "A review of methods to match building energy simulation models to measured data", Renewable and Sustainable Energy Reviews, Volume 37, September 2014, Pages 123-141, ISSN 1364-0321, http://dx.doi.org/10.1016/j.rser.2014.05.007.

Deru, Michael, et al. (2011). U.S. Department of Energy Commercial Reference Building Models of the National Building Stock. National Renewable Energy Laboratory report NREL/TP-5500-46861. Available: http://energy.gov/eere/buildings/commercial-reference-buildings

Garrett, Aaron, New, Joshua R., and Chandler, Theodore (2013). Evolutionary Tuning of Building Models to Monthly Electrical Consumption. Technical paper DE-13-008. In Proceedings of the ASHRAE Annual Conference and ASHRAE Transactions 2013, volume 119, part 2, pp. 89-100, Denver, CO, June 22-26, 2013.

Garrett, Aaron and New, Joshua R. (2013). Trinity Test: Effectiveness of Automatic Tuning for Commercial Building Models. ORNL internal report ORNL/TM-2013/130, March 7, 2013, 24 pages.

Garrett, Aaron and New, Joshua R. (2014). "Scalable Tuning of Building Energy Models to Hourly Data." To appear in the Journal of Energy, 2015.

Judkoff, Ron, Polly, Ben, Marcus Bianchi, Neymark, Joel, and Kennedy, Mike (2011), "Building Energy Simulation Test for Existing Homes (BESTEST-EX): Instructions for Implementing the Test Procedure, Calibration Test Reference Results, and Example Acceptance-Range Criteria. National Renewable Energy Laboratory Report NREL/TP-5500-52414. Available: http://www.nrel.gov/docs/fy11osti/52414.pdf

New, Joshua R., Sanyal, Jibonananda, Bhandari, Mahabir S., Shrestha, Som S. 2012. Autotune E+ Building Energy Models. Proceedings of the 5th National SimBuild of IBPSA-USA, International Building Performance Simulation Association (IBPSA), Aug. 1-3, 2012.

Reddy, T.A., Maor I., Jian, S., and Panjapornporn, C. 2006. Procedures for Reconciling Computer-Calculated Results with Measured Energy Data. ASHRAE Research Project RP-1051, 2006.