Accuracy of HVAC Load Predictions: Validation of EnergyPlus and DOE-2 Using an Instrumented Test Facility

Philip Haves¹, Baptiste Ravache¹, Andres Fergadiotti², Christian Kohler¹, Mehry Yazdanian¹
¹Lawrence Berkeley National Laboratory, Berkeley, United States of America
²Southern California Edison, Rosemead, United States of America

Abstract
Comparisons of predictions from EnergyPlus, DOE-2.1e, and DOE-2.2 with measurements from Lawrence Berkeley National Laboratory’s FLEXLAB® test facility have been conducted with the aim of better understanding the accuracy of these engines. Differences between predicted performance versus actual performance as measured in FLEXLAB, at least for simple buildings, have been identified, starting with key test cases in ASHRAE Standard 140. The averages over the different tests of the magnitudes of the hour-by-hour differences are ~2.5% of the range and the daytime averages, which may be more important for both energy performance and occupant comfort, are ~5%. These results do not support the hypothesis that there is any significant difference in the ability of the three engines to predict thermal loads in spaces with conventional overhead air supply for mixing ventilation.

Introduction
Whole building energy simulation (BES) engines are used to inform the design of utility programs and building energy codes and are increasingly used in the design of commercial buildings. Many analysts and designers and utility program evaluators in the United States use tools based on variants of the DOE-2 engine (Lawrence Berkeley National Laboratory), which was first developed for the US Department of Energy in the 1970s. EnergyPlus™ (U. S. Department of Energy) is the Department of Energy’s free, flagship building energy simulation engine and is intended to replace DOE-2. Comparisons of thermal load predictions made by EnergyPlus and DOE-2.2 (James J. Hirsch et al. a), the simulation engine in eQUEST (James J. Hirsch et al. b), and DOE-2.1e (Lawrence Berkeley National Laboratory), the simulation engine in EnergyPro (EnergySoft), have shown significant differences. Resolution of these differences has an important role to play in informing practitioners, program designers, and other stakeholders regarding the accuracy and robustness of these engines.

Comparisons of predictions from EnergyPlus, DOE-2.1e, and DOE-2.2 with measurements from Lawrence Berkeley National Laboratory’s (LBNL) FLEXLAB® test facility (FLEXLAB® test facility) have been conducted with the aim of better understanding the accuracy of these engines. Seven scenarios (or test cases), covering traditional overhead mixed ventilation designs were tested and each engine predictions showed a consistent level of agreement with the measurements of space level sensible thermal load. These results provide a basis for better understanding the accuracy of these engines in predicting the thermal load elements of whole building energy performance. Detailed measurements of FLEXLAB performance, including indoor temperatures, heat fluxes, and airflow rates in the Heating, Ventilating, and Air Conditioning (HVAC) system, together with hourly weather data and building characteristics, were recorded and used in analyzing the simulation results from the simulation engines.

The results presented here are intended to give both users and practitioners an indication of the accuracy of the thermal load predictions by these BES engines and a way of better understanding and limiting the “risk” associated with these predictions. If the uncertainties in the HVAC system model predictions are relatively small, the load prediction uncertainties will provide a good indication of the overall model algorithm uncertainty in the whole building energy predictions. This paper presents the results of the comparison between the BES models and empirical validation testing, but it also establishes a framework that can be leveraged for validation of other BES by using the same measured data and experimental characteristics.

The paper does not address other sources of uncertainty such as input data errors and user model building errors.

Methods
Laboratory Facility
The experiments were carried out in LBNL’s FLEXLAB test facility in test cells which effectively represent a single perimeter zone in a multistory, relatively deep plan building. Each cell includes its own air handling unit (AHU) with an economizer, hot water, and chilled water coils and supply fan. Each pair of cells has its own chilled water and hot water plant. The thermal isolation resulting from the near adiabatic walls between adjacent cells allows their performance to be analyzed independently.

Experiment Design, Analysis and Instrumentation
For the high priority ASHRAE Standard 140 test cases identified, the simulated building form was modified to one that can be re-created using FLEXLAB and the comparisons repeated to verify that the differences are still observed, even in the more moderate Berkeley climate.
The window and wall sizes and thermal characteristics were modified in simulation to match FLEXLAB. The low-mass cases were reproduced by adding insulation to the top side of the slab-on-grade floor and the high-mass cases were reproduced by exposing the slab. (Each floor slab in FLEXLAB has a continuous layer of rigid insulation separating it from the ground below and hence FLEXLAB is unable to treat slab-on-grade constructions.) The results reported here were obtained using a modified cell configuration with a temporary north wall (partition) that served to isolate the experimental space from the electrical and mechanical equipment located on the north side of the cell. This provided a simpler configuration for modeling using different simulation engines. The modified cell included a well-insulated suspended ceiling and ducted HVAC supply and return. Further details are included in (Haves et al., 2019).

FLEXLAB Modeling
Testbed 3, consisting of Cells 3A and 3B, was modeled using three simulation engines: EnergyPlus, DOE-2.1E, and DOE-2.2. The EnergyPlus model was created using EnergyPlus V8.5 and the OpenStudio® SketchUp plug-in and subsequently upgraded to V8.8. The DOE-2.1e and DOE-2.2 models were created using DrawBDL and eQUEST. DOE-2 uses a one-hour time-step and EnergyPlus was run with a ten-minute timestep.

The simulation input data included:
- Environmental data, including ground temperature, and weather data
- Building geometry
- Building materials and constructions
- Building shading
- Internal gains
- HVAC system and temperature set-points

Environmental Data
Under-floor temperatures were measured using four vertical arrays of sensors, with thermistors spaced at 50 mm (2 in) between the top of the 150 mm (6”) ‘topping slab’ and the bottom of the 150 mm (6”) structural slab, which are separated by a 5” (125 mm) layer of structural polystyrene. The temperatures recorded by the sensors at the bottom of the structural slab were averaged and used as the ground temperature input. The weather data were imported via a weather file created based on measured weather data from the FLEXLAB weather station.

Modeling
The building geometry, materials and construction information used to build the simulation models were obtained from the FLEXLAB as-built drawings.

Floor Modeling
Separate models were produced for the low-mass and high-mass configurations. The only difference between these two configurations was the construction of the floor. In the high-mass configuration, the topping slab was the top layer, with its top surface exposed to the space. In the low-mass configuration, the topping slab was covered with 100 mm (4”) of polyisocyanurate rigid board insulation covered with 13 mm (1/2”) plywood.

Wall Modeling
Whole building simulation engines can only model building components with continuous layers and one-dimensional heat transfer paths. In FLEXLAB, most of the walls, by area, are constructed from structural insulated panels (SIPS) in order to limit 2-D and 3-D heat flow paths to the corners, which also include boxed-in columns. The THERM finite element program was used to identify one-dimensional wall constructions that approximate the thermal performance of two and three-dimensional heat flows in construction elements such as boxed-in columns (Curcija, 2012).

Window Modeling
For some runs, the windows were covered with insulation. Both EnergyPlus and DOE-2 can model moveable insulation on the window that can be controlled with a schedule, allowing one model input file to be used for both insulated and uninsulated periods.

Four construction descriptions were created for the WINDOW 7.4 library: uninsulated single glazing, insulated single glazing, uninsulated window framing and insulated window framing. The appropriate construction description reports for each case were then included in the input files for DOE-2.1e, DOE-2.2 and EnergyPlus.

Internal Gains
The internal heat gains varied diurnally during a number of the experiments, as indicated in Table 1. Annual schedules for internal gains were created, based on the measured electric power consumption.

HVAC System and the Temperature Setpoints
The EnergyPlus model included an Ideal Load System object with a single temperature set-point thermostat assigned to each cell. The DOE-2 model used System=SUM with the proportional thermostat type and a throttling range of 0.1 K. A schedule of zone temperature set-points for the full period of the experiments was created from the average of the 28 air temperature measurements in the cell – four vertical arrays of seven shielded thermistors. This avoids the need for separate runs, each with their own warm-up period, for each of the 12 experiments.

For each experiment, the simulated zone temperature was compared to the measured average dry bulb temperature, which had been used as the zone temperature set-point, and good agreement was found consistently. The FLEXLAB and EnergyPlus ‘zone’ temperatures are dry bulb (air) temperatures. The zone temperature in DOE-2.1e and DOE-2.2 is actually a weighted average of the air and surface temperatures. In practice, the DOE-2 zone temperature is treated as a proxy for the dry bulb temperature, one rationale being that, in multi-zone buildings with lightweight construction, the surface temperatures of the internal partitions and suspended ceilings, at least, tend to follow the zone dry bulb temperature closely enough for most energy calculations and thermal comfort predictions.
Experimental Tests

FLEXLAB tests were performed for the high-mass and low-mass cases. The interior air temperature set-points, internal gain, and window insulation schedules were selected to excite different modes of operation. Sixteen scenarios were defined, consisting of all possible combinations of high-mass vs. low-mass, covered vs. exposed windows, constant or variable internal gains and constant or variable zone set-point temperatures. The cases were structured in order of increasing complexity, starting with gentle transients (solar) to step changes in internal loads and/or zone temperature set-points. In several variable set-point cases, there was a substantial step change (5-10 K) in the zone air temperature set-point twice per day. The intention was to test the transient response of the simulation engines. However, in some cases, there was overshoot in the FLEXLAB zone air temperature control loop, which increased the higher frequency components of the response.

Tests were performed for six of these scenarios, selected to address different combinations of variable envelope conduction, window solar heat gain and charging/discharging of thermal mass, as shown in Table 1. The rightmost column indicates the closest corresponding BESTEST; the BESTEST numbering scheme has been extended to indicate new variants in which the internal gains were varied diurnally.

The tests involved defining schedules of zone air temperature set-points and lighting and miscellaneous equipment switching in the control system. Test periods were typically three to five days, though, in some cases, operational problems resulted in the measurements for some part of the period being rejected.

The primary output for each test was a set of one-minute measurements for estimating the load: supply and return air temperatures, measured at the supply diffusers and return grill, together with measurements of the supply air flow rate. Each of the temperature sensors in each cell – one in each of the four supply diffusers and three in the single return grill – was carefully calibrated and in-situ cross-calibration was then performed to minimize the uncertainty in the supply-return temperature difference, which is estimated to be ~0.15 K (0.27 F). The airflow station in the supply duct was calibrated against two different reference airflow measurement systems and the resulting uncertainty is estimated to be ≤3% across the range of flow rates used in the tests. When the supply-return temperature difference is ~5 K, the uncertainty in the measurement of thermal load is then ~5%.

Zone air temperature, surface temperature, and surface heat flux measurements were also collected. Each sensor was sampled once per second and one-minute averages were archived for use in visualization and analysis. The initial intention was to use water-side measurements of the duties of the heating and cooling coils, together with the measured fan power, to determine the rate at which heat was added to, or removed from, the main test zone. However, the following sources of uncertainty resulted in the decision to use a direct air-side measurement:

<table>
<thead>
<tr>
<th>Scenario : Test</th>
<th>Therm Mass</th>
<th>Window Insul</th>
<th>Internal Gain (W)</th>
<th>Zone Temperature</th>
<th>Fan Speed (%)</th>
<th>Test Dates</th>
<th>Best-Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>3:01</td>
<td>Low</td>
<td>No</td>
<td>0</td>
<td>22 C / 22 C</td>
<td>25</td>
<td>Jul22-Jul26, 2017</td>
<td>600</td>
</tr>
<tr>
<td>3:02</td>
<td>Low</td>
<td>No</td>
<td>0</td>
<td>22 C / 22 C</td>
<td>40</td>
<td>Feb3-Feb5, 2018</td>
<td>600</td>
</tr>
<tr>
<td>3:03</td>
<td>Low</td>
<td>No</td>
<td>0</td>
<td>22 C / 22 C</td>
<td>Varies 10-50</td>
<td>Feb7-Feb9, 2018</td>
<td>600</td>
</tr>
<tr>
<td>3:04</td>
<td>Low</td>
<td>No</td>
<td>0</td>
<td>22 C / 22 C</td>
<td>50:8-18, 25:18-8</td>
<td>Feb12-Feb18, 2018</td>
<td>600</td>
</tr>
<tr>
<td>3:05</td>
<td>Low</td>
<td>No</td>
<td>0</td>
<td>22 C / 22 C</td>
<td>25</td>
<td>Feb20-Feb23, 2018</td>
<td>600</td>
</tr>
<tr>
<td>6:01</td>
<td>High</td>
<td>Yes</td>
<td>0/1440 W</td>
<td>25 C / 25 C</td>
<td>50→40→50</td>
<td>Jul1-Jul3, 2017</td>
<td>800VIG</td>
</tr>
<tr>
<td>11:01</td>
<td>Low</td>
<td>No</td>
<td>0</td>
<td>30 C / 20 C</td>
<td>50</td>
<td>Nov12-Nov16, 2016</td>
<td>640</td>
</tr>
<tr>
<td>11:02</td>
<td>Low</td>
<td>No</td>
<td>0</td>
<td>30 C / 20 C</td>
<td>50</td>
<td>Nov24-Nov28, 2016</td>
<td>640</td>
</tr>
<tr>
<td>11:03</td>
<td>Low</td>
<td>No</td>
<td>0</td>
<td>30 C / 20 C</td>
<td>50</td>
<td>Nov30-Dec03, 2016</td>
<td>640</td>
</tr>
<tr>
<td>12:01</td>
<td>High</td>
<td>No</td>
<td>0</td>
<td>26 C / 22 C</td>
<td>25</td>
<td>Jul10-Jul14, 2017</td>
<td>940</td>
</tr>
<tr>
<td>14:01</td>
<td>High</td>
<td>Yes</td>
<td>0/1440 W</td>
<td>22 C / 26 C</td>
<td>25</td>
<td>Jul6-Jul10, 2017</td>
<td>840VIG</td>
</tr>
<tr>
<td>15:01</td>
<td>Low</td>
<td>No</td>
<td>0/435 W</td>
<td>30 C / 20 C</td>
<td>25</td>
<td>Dec15-Dec17, 2016</td>
<td>640VIG</td>
</tr>
</tbody>
</table>
• **Fan power:** Direct measurement of the power supplied to the fan motor by the variable frequency drive (VFD) is uncertain/expensive because the harmonic content of the current. The VFD can be bypassed, but only at one fan speed; different fan speeds were used in the tests to minimize the uncertainty in the measurement of heating/cooling rate while maintaining adequate heating/cooling capacity.

• **Heat gains/losses:** Heat gains/losses through the (insulated) walls of the supply and return ducts that are significant enough to be very difficult to correct sufficiently, given the spatial and temporal variations in the temperature adjacent to the exterior surfaces of the ducts.

### Results
Detailed comparisons of the measurements with the outputs from EnergyPlus, DOE-2.2, and DOE-2.1e have been performed, and results are shown in this next section. The primary comparison is in terms of the thermal load, measured in FLEXLAB by performing an air-side heat balance on the test cell, as discussed above.

#### Comparisons Between EnergyPlus, DOE-2.2, DOE-2.1E and FLEXLAB

Figure 1 shows an example of the standard set of plots and a summary table used to facilitate the analysis used to generate the results presented here. A comparison between the simulation outputs (heat removal rate) from each of the three engines is presented; first as a time-series and then in cumulative/integrated form. The band around the measured heat removal rate line represents an estimate of the effect of the uncertainties in the supply-return air temperature difference and the air flow rate. The numerical values at the right-hand end of the cumulative plot can be used to obtain the cumulative bias error (MBE) by subtracting the simulation engine value from the FLEXLAB value.

Table 2 presents two different but complementary metrics that characterize the case-by-case differences between the FLEXLAB measurements of heat removal rate and the predictions of DOE-2.1e, DOE-2.2 and EnergyPlus. The first metric is the Range-Normalized Mean Bias Error (RNMBE). The differences between hourly samples (\(i\)), of the simulation prediction (\(s\)), and the FLEXLAB measured value (\(f\)), are averaged for each test and normalized to the range of the FLEXLAB measurements for that test, as defined in Equation 1.

\[
RNMBE = \frac{1}{\max(f) - \min(f)} \frac{\sum_{i=1}^{n} \left| s(i) - f(i) \right|}{n}
\]  

(1)

This is in contrast to the mean percentage error (MPE), where the normalization in the denominator varies from sample to sample, thereby giving a different weight to each sample. Another consideration is the quantity to be used in the denominator; common choices are the mean value of \(f\), the mean absolute value of \(f\) and the range of \(f\). Since \(f\) can be either positive (cooling load) or negative (heating load), its mean value may be inappropriately small. The range was selected in preference to the mean absolute value since it better characterizes the simulation application. Since the mean of the actual differences, rather than the absolute differences, is calculated, fluctuations cancel out, yielding the underlying trend, or bias, for that test.

The second metric is the Range-Normalized Root Mean Square Error (RNRMSE). The square root of the average of the square of the differences between hourly samples of the simulation prediction and the FLEXLAB measured value is normalized to the range of the FLEXLAB measurements for that test, as defined in Equation 2. Range normalization was selected for the reason described above for RNMBE.

\[
RNRMSE = \frac{1}{\max(f) - \min(f)} \sqrt{\frac{\sum_{i=1}^{n} \left( s(i) - f(i) \right)^2}{n}}
\]  

(2)

Fluctuations are accentuated; there is no cancellation of positive and negative errors and squaring the error gives more weight to outliers, which tend to occur more frequently during transients.

Normalized Mean Bias Error (strictly, Mean Bias) and the Coefficient of Variation of the Root Mean Square Error, both of which are normalized to the mean value rather than the range, are statistical metrics that have been adopted for use in the International Performance Measurement and Verification Protocol – IPMVP – (Efficiency Valuation Organization) and by ASHRAE Guideline 14P Measurement of Energy, Demand, and Water Savings (ASHRAE, 2014) to characterize how well a simulation model has been calibrated to measured data prior to use in a Measurement and Verification (M&V) exercise. This is a different situation to that of the work presented here, which does not involve calibrated simulation and is concerned with thermal load, which may be positive or negative, rather than with energy consumption. In most situations, a simulation model constructed from as-built drawings and other design information is not expected to be as accurate as a model that is also calibrated with real measurements.

In effect, the purpose of the work reported here is to determine how accurate a carefully constructed simulation model is likely to be before any calibration, such as in the context of design for new construction. In particular, it is not appropriate to apply the acceptance criteria used in IPMVP and Guideline 14 for the adequacy of simulation model calibration to determine the adequacy of simulation engines in the context of the work presented here, since no calibration is involved. Three other confounding factors are:

- The operating conditions in the tests reported here were deliberately designed to stress the engines. (For example, there was no dead-band between the heating and cooling set-points for the zone temperature.)
- The tests reported only address the thermal load in an occupied space and do not consider the performance of the HVAC system(s).
Scenario 3:01: Low-mass, Exposed Windows, No Internal Gains, Constant Set-point (22°C)

Figure 1: Graphical representation of an example test

Table 2: Tabular summary of the example test illustrated in Figure 1

<table>
<thead>
<tr>
<th></th>
<th>Day</th>
<th>Night</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>RNMBE</td>
<td>-5.52%</td>
<td>-16.40%</td>
<td>-18.50%</td>
<td>0.74%</td>
</tr>
<tr>
<td>RNRMSE</td>
<td>8.18%</td>
<td>19.20%</td>
<td>21.20%</td>
<td>3.18%</td>
</tr>
</tbody>
</table>

The decision to normalize the calculated values of the Mean Bias Error (MBE) and the Root Mean Square Error (RMSE) to the difference between the maximum and the minimum value of the load - cooling and heating, was made to compare the uncertainty to the full range of thermal loads that the building is predicted to encounter in a particular run. The corresponding metrics used in IPMVP and Guideline 14 are normalized to the mean absolute value and so are not directly comparable to the RNMBE and RNRMSE metrics used here.

These patterns may be consistent between day and night or may be different between day and night. Better agreement is often observed at night, presumably because of the absence of solar radiation as a complicating factor.
Data Analysis

Table 3 show the values of the range-normalized mean bias error (RNMBE) and the range-normalized root mean square error (RNRMSE) between the predictions of each engine and the FLEXLAB measurements, aggregated over daytime and night-time, for each of the 12 cases. The average of the RNMBE values for all the cases provides an indication of the overall bias, with cancellation between over-prediction and under-prediction of different cases. The Absolute Average values are averages of the absolute values of the RNMBE for each case and so there is no cancellation of over-prediction and under-prediction between the different cases, making these values a better indication of the uncertainty in any particular run and, by extension, in a set of runs of the same model as part of a design project. The uncertainties in the tables are standard deviations of the populations.

The average of the RNRMSE values for all the cases provides an indication of the variability of the error for all the cases. Because the value of RNRMSE is always positive for each run, there is no cancellation when the values for the different cases are averaged. The relatively large values of RNRMSE – averaging ~10% - indicates that there can be significant variation at the hourly level. Some of the larger values may not be representative of real operation, in that large simultaneous step changes in the internal gain and the zone temperature set-point were made in some runs to create ‘stress cases’.

Some significant differences were observed between the FLEXLAB measurements and the predictions of each simulation engine; the test runs can be categorized as follows:

- Good agreement (RNMBE < 5%, RNRMSE < 10%) between all the engines and FLEXLAB (Cases 3:02, 3:04, 11:02, 11:03, 15:01)
- Good agreement (RNMBE < 5%, RNRMSE < 10%) between all the engines and FLEXLAB (Cases 3:02, 3:04, 11:02, 11:03, 15:01)
- Good agreement between EnergyPlus and FLEXLAB but poorer agreement between the DOE-2 engines and FLEXLAB (Cases 3:01, 3:03)
- Better agreement between FLEXLAB and one of the two DOE-2 engines than with the other DOE-2 engine and with EnergyPlus (Cases 6:01, 14:01)
- Poorer agreement between FLEXLAB, EnergyPlus and the DOE-2 engines (Cases 3:05, 12:01)

Discussion

Both the steady state/low frequency deviations between the engine predictions and FLEXLAB measurement, as indicated by RNMBE, and the high frequency deviations, as indicated by RNRMSE, vary significantly from test to test. In general, experimental results suggest that any systematic bias in the engines is quite small; EnergyPlus exhibits a barely significant bias of 2.3±2.1% towards overprediction of cooling load, possibly with an element of underprediction of heating load. The corresponding values for the DOE-2 engines are not significant. The quasi-steady-state variability from run-to-run for EnergyPlus, as indicated by the average of the absolute values of the bias errors (RNMABE) is only slightly

<table>
<thead>
<tr>
<th>SCENARIO</th>
<th>RNMBE</th>
<th>RNRMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>EPLUSTOMEAS</td>
<td>DOE21TOMEAS</td>
</tr>
<tr>
<td>RUN</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3:01</td>
<td>-1.87%</td>
<td>-8.72%</td>
</tr>
<tr>
<td>3:02</td>
<td>4.01%</td>
<td>3.31%</td>
</tr>
<tr>
<td>3:03</td>
<td>4.45%</td>
<td>3.93%</td>
</tr>
<tr>
<td>3:04</td>
<td>4.23%</td>
<td>2.53%</td>
</tr>
<tr>
<td>3:05</td>
<td>6.69%</td>
<td>4.20%</td>
</tr>
<tr>
<td>6:01</td>
<td>1.61%</td>
<td>-1.28%</td>
</tr>
<tr>
<td>11:01</td>
<td>2.59%</td>
<td>0.49%</td>
</tr>
<tr>
<td>11:02</td>
<td>1.18%</td>
<td>-1.17%</td>
</tr>
<tr>
<td>11:03</td>
<td>1.46%</td>
<td>-0.65%</td>
</tr>
<tr>
<td>12:01</td>
<td>1.86%</td>
<td>-0.44%</td>
</tr>
<tr>
<td>14:01</td>
<td>0.81%</td>
<td>-0.64%</td>
</tr>
<tr>
<td>15:01</td>
<td>1.04%</td>
<td>-1.79%</td>
</tr>
<tr>
<td>Average</td>
<td>2.3±2.1%</td>
<td>0.0±3.3%</td>
</tr>
<tr>
<td>Absolute Average</td>
<td>2.6±1.7%</td>
<td>2.4±2.3%</td>
</tr>
</tbody>
</table>
significant (2.6±1.7%), while corresponding values for the DOE-2 engines are barely significant.

The response to transients, as indicated by the root-mean-square error (RNRMSE), is also quite similar for all three engines, i.e. 8-10%. This is despite the use of a relatively short time-step (10 min) in EnergyPlus and a one-hour time-step in DOE-2.1e and DOE-2.2. In contrast to the mean bias errors, these values are all greater than three standard deviations, suggesting that the errors in the transient responses of the engines are more significant than the quasi-static errors.

The experimental results indicate that the average value of RNRMSE for all engines and all tests is -4.4% during the day and 2.7% during the night and the corresponding values of RNRMSE are 13.7% and 8.5%; these differences indicate that there are greater bias errors, and greater transient deviations, during the day. It seems probable that these are caused by solar gain, though the number of insulated window tests is not great enough to allow solar effects to be separated from, for example, the effect of other transients occupying a larger fraction of the shorter time interval (10 hours vs. 14 hours).

The EnergyPlus daytime average bias error of 0.3% is substantially smaller than the daytime average absolute bias error of 4.7%, indicating substantial cancelation between positive and negative bias errors for different cases during the daytime. By contrast, the daytime average bias error for the DOE-2 engines is -3.7%, which is not much smaller than the daytime average absolute bias error of 5.4%, indicating only modest cancelation between positive and negative bias errors for different cases during the daytime. Conversely, the night-time average bias errors fall in the range 2.6% - 3.8%, only slightly smaller than the night-time average absolute bias errors of 3.5% - 4.1%, indicating modest cancelation between positive and negative bias errors for different cases during the night-time.

The practical significance of the difference between the daytime and night-time results is that both energy consumption and the importance of thermal comfort are generally greater during the daytime, giving greater significance to the daytime results and diminishing the importance of the 24-hour average results.

Suggestions of possible causes of the differences between the predictions of the simulation engines and the measurements were made by members of the project Technical Advisory Group; these included limitations of the models of long wave infra-red radiation from the sky. EnergyPlus uses a model developed by Clark and Allen (1976), which is based on observations of a narrower range of climates than the Berdahl and Martin (1984) model used in DOE-2. A sensitivity test was performed using EnergyPlus in which the emissivity of the roof was changed from the default value of 0.9 to zero and no perceptible difference in the nocturnal heating load was observed, confirming that the ~350 mm (14 in.) of polyurethane and polystyrene insulation in the roof of FLEXLAB provides effective isolation of the ceiling void from the ambient conditions. That said, the different sky models will predict different values of long wave irradiance incident on the window, and this may account for some part of the modest differences between the engines in their predictions of night-time thermal load observed in some of the cases with uncovered windows. The default window algorithm in DOE-2 has some differences to the EnergyPlus window algorithm selected for the analysis reported here; further analysis is required.

In the initial stages of the analysis of the measurements reported here, it was observed that, in terms of thermal load, the response of EnergyPlus to substantial step changes in the internal heat gain or the zone temperature set-point appeared underdamped compared to the transient responses of FLEXLAB and the DOE-2 engines. Investigation determined that this effect was driven by the use of the “CeilingDiffuser” interior convection algorithm in EnergyPlus, rather than the default “TARP” interior convection model. Under the range of conditions prevailing in most of the tests, the CeilingDiffuser model predicts significantly greater values of the interior surface convection coefficient than does the TARP model or the default fixed values used in DOE-2, resulting in better agreement with the measurements, except in 12:01. The effect of the choice of interior convection algorithm in EnergyPlus is explored in more detail in a companion paper (Li et al. 2019).

The TARP model was used to generate the EnergyPlus results presented in this report and response can be seen to be overdamped relative to FLEXLAB in some cases (For example, 12:01, though not in others). Trials were made using the Beausoleil-Morrision hybrid model (Beausoleil-Morrision, 2000) included as an option in EnergyPlus, which combines the momentum-driven CeilingDiffuser model with the buoyancy-driven TARP model, but no reduction in the differences between the EnergyPlus and the FLEXLAB transient responses was observed. At the time of writing, possible refinements to the CeilingDiffuser model that may help to resolve the differences with the FLEXLAB results were being investigated by the EnergyPlus development team.

Not enough data were obtained for the high-mass cases to be able to separate the effect of high-mass and low-mass. In particular, it is not possible to assert that the high-mass cases have higher values of RNRMSE, though the data are not inconsistent with the high-mass cases having higher values of RNRMSE. Additional high-mass runs are required for more definitive statements to be made.

A simplification adopted in the work reported here was the use of a single variable, heat extraction rate, which is positive when cooling and negative when heating. The main limitation of this simplification is that there can be cancellation of differences in the magnitude of the heat rate between simulation and measurement when there is significant heating and significant cooling during a single experiment. However, substantial heating (>20% of the cooling load) occurred in only three of the twelve experiments used in the analysis (12:01, 14:01 and 15:01) and examination of the data does not indicate that heating/cooling cancelation has a significant effect on the
results of the analysis – this question is explored in the companion paper by Li et al. (2019).

**Conclusion**

The differences between the thermal load predictions of EnergyPlus and the DOE-2 engines and the FLEXLAB measurements are generally relatively small for the conditions in the tests described in this report. The averages over the different tests of the magnitudes of the hour-by-hour differences are ~2.5% of the range. The daytime averages, which may be more important for both energy performance and occupant comfort, are ~5%, presumably because of the greater complexity involved in simulating in the presence of solar radiation.

The 24-hour and daytime systematic bias errors are in the range ~4.2% – 2.3%; in each case, the corresponding standard deviation exceeds the average mean bias error, so it is not possible to make statements about the magnitudes, or even the signs, of the average mean bias errors, particularly since the individual values do not appear to be normally distributed.

EnergyPlus typically over predicts the cooling load and/or underpredicts the heating load by ~1.5% and the DOE-2 engines typically under predict the cooling load by approximately the same amount. The Root Mean Square Error is relatively more sensitive to shorter term variations in the difference between predicted and measured loads; the three engines have fairly similar values, ~10%.

The implication of these results is that users, both designers and efficiency incentive program analysts, can use either EnergyPlus or DOE-2 to model the heating and sensible cooling loads in conventional commercial buildings equipped with overhead mixing ventilation with a similar level of confidence while limiting the “risk” associated with these predictions.

Given the limited nature of the cases studied, these results do not support the hypothesis that there is any significant difference in the ability of the three engines to predict thermal loads in spaces with conventional overhead air supply for mixing ventilation. The results of this study, and the conclusions stated above, may not be fully applicable to spaces with furniture or sources of internal heat gain that have different characteristics to those used in the study. The results, and conclusions, do not apply to other methods of space conditioning, such as radiant heating and cooling, UFAD, displacement ventilation, and natural ventilation.

**Acknowledgements**

The work reported here was supported by California utility customers under the auspices of the California Public Utilities Commission and by the Assistant Secretary for Energy Efficiency and Renewable Energy, Office of Building Technology, State and Community Programs of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. The authors wish to thank Darryl Dickeroff, Christian Fitting and Ari Harding (LBNL) for their assistance with the experimental measurements, Zhan Yang (University of Southern California) for assistance with the compilation of construction details and Veronica Soebarto (University of Adelaide, Australia) for assistance with the analysis and presentation of the results.

**References**


